

EIN ALGORITHMUS FÜR DIE TSCHEBYSCHJEFF-LÖSUNG VON LINEAREN ÜBERBESTIMMTEN GLEICHUNGSSYSTEMEN

INAUGURAL-DISSERTATION

zur Erlangung
der philosophischen Doktorwürde

vorgelegt der
Philosophischen Fakultät II
der Universität Zürich

von
ETEOKLIS REKKAS
aus Athen (Griechenland)

Begutachtet von Herrn Prof. Dr. H. R. Schwarz

Zürich 1981
Zentralstelle der Studentenschaft

EIN ALGORITHMUS FÜR DIE TSCHEBYSCHJEFF-LÖSUNG VON LINEAREN ÜBERBESTIMMTEN GLEICHUNGSSYSTEMEN

INAUGURAL-DISSERTATION

zur Erlangung
der philosophischen Doktorwürde

vorgelegt der
Philosophischen Fakultät II
der Universität Zürich

von
ETEOKLIS REKKAS
aus Athen (Griechenland)

Begutachtet von Herrn Prof. Dr. H. R. Schwarz

Zürich 1981
Zentralstelle der Studentenschaft



Inhalt

Bezeichnungen

Einführung

I. Allgemeine Begriffe und Sätze der numerischen linearen Algebra

Normen, Matrizen	1
Unitäre Zerlegung	6
Aenderungsformel bei symmetrischen Matrizen	6
Orthogonalisierungsproblem	7
Pseudoinverse Matrix	8
Rundungsfehler bei Gleitpunktrechnung	11

II. Uebersicht über das Tschebyscheff-Ausgleichsproblem

Problemstellung	13
Theoretische Grundlagen	14
Austauschsatz	15
Ausgleichung nach kleinsten Quadraten	16

III. Aenderungsformel der pseudoinversen Matrix A_I^+ .

Lösungsformeln

Bestimmung des Vektors $\vec{f}_I$	19
Aenderungsformel der Matrix A_I^+	23
Lösung der Referenz $[I']$ nach der Methode der kleinsten Quadrate	29
Bestimmung der Referenzabweichung	30
Austauschkriterium	31

IV. Bestimmung der maximalen Referenzabweichung bei jedem Schritt. Stabilität der Aenderungsformel.

Beispiele

Bestimmung der maximalen Referenzabweichung	34
Kriterium	44
Fehlerabschätzung der Matrix A_I^+	45
Iterative Verbesserung der Matrix A_I^+	48
Beschreibung des Algorithmus	50
Beispiele	52
Literatur	59



Digitized by the Internet Archive
in 2026

BEZEICHNUNGEN

SYMBOL

BEDEUTUNG

$\mathbb{R}$

die Menge der reellen Zahlen.

$\mathbb{R}^n$

n-dimensionaler euklidischer Raum.

ϵ

Abkürzung für "Element von".

$$\operatorname{sgn}(\alpha) = \begin{cases} -1 \\ +1 \end{cases}$$

für $\alpha \in \mathbb{R}, \alpha < 0$

für $\alpha \in \mathbb{R}, \alpha \geq 0$

$$|\alpha| = \begin{cases} -\alpha \\ +\alpha \end{cases}$$

für $\alpha \in \mathbb{R}, \alpha < 0$

für $\alpha \in \mathbb{R}, \alpha \geq 0$

$$\vec{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$$

Spaltenvektor der Dimension n.

$\langle \cdot, \cdot \rangle$

Skalarprodukt.

A^T

die zu A transponierte Matrix.

I_n

die (n x n)-Einheitsmatrix.

$\operatorname{sp}(A)$

Spur der Matrix A.

$\rho(A)$

Spektrum der Matrix A.

$o(I)$

Anzahl Elemente der Menge I .

$$I_0/I = \{i | i \in I_0 \text{ und } i \notin I\}$$

EINFUEHRUNG

Ueberbestimmtes Gleichungssystem wird ein lineares Gleichungssystem mit mehr Gleichungen als Unbekannten genannt. Es stellt sich das Problem, ein solches System, wenn schon nicht exakt, so doch möglichst gut zu lösen. Als möglichst gute Lösungen von solchen Systemen kennt man unter anderem die Minimierung der euklidischen Norm des Fehlers und die Minimierung der Maximumnorm des Fehlers.

Im Fall der Minimierung der euklidischen Norm hat man ein Problem der Ausgleichsrechnung im engeren Sinne, das bereits von Gauss ("Methode der kleinsten Quadrate") studiert wurde. Für eine ausführlichere Behandlung der Methode der kleinsten Quadrate sei auf [15,27] verwiesen.

Bei der Minimierung der Maximumnorm ("Tschebyscheff-Norm") des Fehlers eines linearen überbestimmten Gleichungssystems hat man mit einem Tschebyscheff-Ausgleichsproblem, kurz T.A.P., zu tun. In diesem Fall hat Stiefel [29] für die Berechnung der besten Lösung das Austauschverfahren entwickelt. Die Herleitung der Rechenregel in Stiefels Verfahren [29] ist an die Voraussetzung gebunden, dass die homogenen Teile von je n linearen Gleichungen (n = Anzahl Unbekannten) linear unabhängig sind ("Haarsche-Bedingung").

In [20,30] wird festgestellt, dass Stiefels Austauschverfahren zur dualen Simplexmethode, angewendet auf ein zugeordnetes lineares Programmierungsproblem, äquivalent ist.

Diese Äquivalenz wird auch in [4] betrachtet, wo modifizierte Austauschregeln entwickelt werden, um den Fall zu erfassen, bei dem die Haarsche-Bedingung nicht erfüllt ist.

Die Algorithmen von [1,24] liefern die Lösung des T.A.P., indem dieses Problem ebenfalls auf eine lineare Optimierungsaufgabe zurückgeführt wird, wobei [1] gewisse Modifikationen in das duale lineare Optimierungsproblem einführt, so dass die Lösung des T.A.P. mit weniger Schritten erreicht wird. Die Algorithmen [1,24] sind anwendbar, wenn die Koeffizientenmatrix einen Rang $< n$ hat.

In der Arbeit von Cheney [8] werden lineare überbestimmte Gleichungssysteme ausführlich behandelt. In der gleichen Arbeit findet man sechs Algorithmen zur Berechnung der Tschebyscheff-Lösung von solchen Systemen. Die Algorithmen I und III liefern die Tschebyscheff-Lösung linearer überbestimmter Gleichungssysteme mit unendlich vielen Gleichungen und die Algorithmen IV, VI die Tschebyscheff-Lösung linearer überbestimmter Gleichungssysteme mit Nebenbedingungen.

In den Arbeiten [10,17] werden Relationen zwischen der Tschebyscheff-Lösung eines linearen Gleichungssystems aus $n+1$ Gleichungen in n Unbekannten mit der Lösung des gleichen Problems nach der Methode der kleinsten Quadrate hergeleitet, wobei in Duris Arbeit [10] ein Algorithmus beschrieben wird, der die Tschebyscheff-Lösung eines linearen überbestimmten Gleichungssystems mit Hilfe solcher Relationen berechnet.

Weitere Algorithmen, die das T.A.P. lösen, findet man in [3,9,11,13,16,18,22,33].

In der vorliegenden Arbeit wird ein stabiles Verfahren beschrieben, welches die Tschebyscheff-Lösung eines linearen überbestimmten Gleichungssystems mit Hilfe der Methode der kleinsten Quadrate in weniger Schritten löst.

In Kapitel I der vorliegenden Arbeit werden die benötigten Definitionen und Sätze erwähnt.

In Kapitel II wird das Austauschverfahren skizziert, das uns ermöglicht, die Tschebyscheff-Lösung eines linearen überbestimmten Gleichungssystems zu berechnen. Weiter wird auf den Zusammenhang zwischen der Tschebyscheff-Lösung einer Referenz und der Lösung, die man nach der Methode der kleinsten Quadrate der gleichen Referenz erhält, hingewiesen.

In Kapitel III wird untersucht, wie bei der Aenderung einer Zeile einer Referenzmatrix die entsprechende Pseudoinverse sich ändert, sowie eine Formel hergeleitet, die die Lösung einer neuen Referenz nach der Methode der kleinsten Quadrate angibt, wenn die Lösung der alten Referenz bekannt ist.

In Kapitel IV wird ein Kriterium hergeleitet, das uns die maximale Referenzabweichung bei jedem Schritt ergibt. Ferner wird die Stabilität der vorgeschlagenen Aenderungsformel für die Pseudoinverse untersucht. Zum Schluss werden numerische Beispiele behandelt, die mit dem vorgeschlagenen Algorithmus und den Algorithmen [8] und [1] am Rechenzentrum der Universität Zürich gelöst worden sind.

I. ALLGEMEINE BEGRIFFE UND SAETZE DER NUMERISCHEN LINEAREN ALGEBRA.

In diesem Abschnitt werden in dieser Arbeit benötigte Definitionen und Sätze festgehalten.

I.1. Definition.

Die Abbildung $\| \cdot \| : \mathbb{R}^n \rightarrow \mathbb{R}$ heisst Vektornorm, wenn folgende Eigenschaften erfüllt sind.

$$\| \vec{x} \| \geq 0 \quad \vec{x} \in \mathbb{R}^n, \text{ " = " genau dann, wenn } \vec{x} = 0 \quad (\text{I.1.1})$$

$$\| \alpha \vec{x} \| = |\alpha| \| \vec{x} \| \quad \alpha \in \mathbb{R}, \vec{x} \in \mathbb{R}^n \quad (\text{I.1.2})$$

$$\| \vec{x} + \vec{y} \| \leq \| \vec{x} \| + \| \vec{y} \| \quad \vec{x}, \vec{y} \in \mathbb{R}^n. \quad (\text{I.1.3})$$

Wir benützen im folgenden nur die Normen:

$$\| \vec{x} \|_1 = \sum_{i=1}^n |x_i| \quad (\text{I.1.4})$$

$$\| \vec{x} \|_2 = (\langle \vec{x}, \vec{x} \rangle)^{\frac{1}{2}} \quad (\text{I.1.5})$$

$$\| \vec{x} \|_{\infty} = \max_{1 \leq i \leq n} |x_i|. \quad (\text{I.1.6})$$

I.2. Definition.

Die Abbildung $\| \cdot \| : \{A = (a_{ij}) \mid a_{ij} \in \mathbb{R}\} \rightarrow \mathbb{R}$ heisst Matrixnorm, wenn folgende Bedingungen erfüllt sind:

$$\| A \| \geq 0 \quad \text{und} \quad \| A \| = 0 \text{ nur für } A = 0 \quad (\text{I.2.1})$$

$$\| cA \| = |c| \| A \| \quad \text{für } c \in \mathbb{R} \quad (\text{I.2.2})$$

$$\| A+B \| \leq \| A \| + \| B \| \quad (\text{Dreiecksungleichung}) \quad (\text{I.2.3})$$

$$\| A \cdot B \| \leq \| A \| \cdot \| B \| \quad (\text{I.2.4})$$

Zu jeder Vektornorm kann man eine zugeordnete Matrixnorm durch

$$\|A\| = \sup_{\vec{x} \neq 0} \frac{\|A\vec{x}\|}{\|\vec{x}\|} \quad (1.2.5)$$

definieren.

Den drei Vektornormen (I.1.4–I.1.6) entsprechen die folgenden zugeordneten Matrixnormen:

$$\|A\|_1 = \max_j \sum_i |a_{ij}| \quad (1.2.6)$$

$$\|A\|_2 = (\text{maximaler Eigenwert von } A^T A)^{\frac{1}{2}} \quad (1.2.7)$$

$$\|A\|_\infty = \max_i \sum_j |a_{ij}|. \quad (1.2.8)$$

I.3. Definition.

Eine Matrixnorm $\|A\|$ heisst kompatibel mit einer Vektornorm $\|\vec{x}\|$, falls für jeden beliebigen Vektor $\vec{x}$ die Relation

$$\|A\vec{x}\| \leq \|A\| \cdot \|\vec{x}\| \quad (1.3.1)$$

erfüllt ist.

Die Matrixnormen (I.2.6–I.2.8) sind mit den entsprechenden Vektornormen kompatibel.

Zur Vektornorm $\|\vec{x}\|_2$ gibt es auch noch andere kompatible Matrixnormen, beispielsweise die Frobenius-Norm

$$\|A\|_F = \left(\sum_i \sum_j (a_{ij})^2 \right)^{\frac{1}{2}}. \quad (1.3.2)$$

I.4. Satz.

Es gelten die folgenden Abschätzungen:

$$\|\vec{x}\|_\infty \leq \|\vec{x}\|_2 \leq n^{\frac{1}{2}} \|\vec{x}\|_\infty \quad (1.4.1)$$

$$\|\vec{x}\|_\infty \leq \|\vec{x}\|_1 \leq n \|\vec{x}\|_\infty \quad (1.4.2)$$

$$\|A\|_2 \leq (\|A\|_1 \|A\|_\infty)^{\frac{1}{2}} \quad (1.4.3)$$

$$\|A\|_2 \leq \|A\|_F = \| |A| \| \leq n^{\frac{1}{2}} \|A\|_2. \quad (1.4.4)$$

Für die Definitionen I.1–I.3, sowie für den Beweis des Satzes I.4. siehe [27, 32].

I.5. Unitäre Zerlegung.

A sei eine reelle $(m \times n)$ -Matrix mit $m > n$ und $\text{Rang}(A) = n$. Dann gibt es eine spaltenorthonormierte $(m \times n)$ -Matrix U und eine orthogonale $(n \times n)$ -Matrix V, so dass

$$U^T A V = S \quad \text{d.h.} \quad A = U S V^T \quad (I.5.1)$$

gilt.

Dabei ist

$$S = \begin{bmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ 0 & & \sigma_n & \\ & & & \ddots \end{bmatrix}$$

und es gilt:

$$U^T U = V^T V = V V^T = I_n \quad (I.5.2)$$

$$A \vec{v}_i = \sigma_i \vec{u}_i \quad (i=1, \dots, n) \quad (I.5.3)$$

$$A^T \vec{u}_i = \sigma_i \vec{v}_i \quad (i=1, \dots, n), \quad (I.5.4)$$

wobei $\sigma_1 > \sigma_2 > \dots > \sigma_n$ die singulären Werte der Matrix A sind und $\vec{u}_i, \vec{v}_i$ ($i=1, \dots, n$) die Spaltenvektoren der Matrix U bzw. V sind [12, 15].

I.6. Aenderungsformel bei symmetrischen Matrizen.

Für jede nichtsinguläre symmetrische $(n \times n)$ -Matrix C und alle Vektoren $\vec{x}, \vec{z} \in \mathbb{R}^n$ gelten:

$$(C + \vec{x} \vec{z}^T)^{-1} = C^{-1} - \frac{C^{-1} \vec{x} \vec{z}^T C^{-1}}{1 + \vec{z}^T C^{-1} \vec{x}} \quad (I.6.1)$$

$$(C - \vec{x} \vec{z}^T)^{-1} = C^{-1} + \frac{C^{-1} \vec{x} \vec{z}^T C^{-1}}{1 - \vec{z}^T C^{-1} \vec{x}} \quad (I.6.2)$$

Ist $\vec{z}^T C^{-1} \vec{x} = -1$, so ist $(C + \vec{x} \vec{z}^T)$ singulär.

Ist $\vec{z}^T C^{-1} \vec{x} = 1$, so ist $(C - \vec{x} \vec{z}^T)$ singulär [10, 34].

I.7. Das Orthogonalisierungsproblem.

Gegeben sei eine Menge von n linear unabhängigen Vektoren

$\vec{a}_i$ ($i=1, \dots, n$) eines m -dimensionalen Vektorraumes

$\{\vec{a}_i | i=1, \dots, n, \vec{a}_i \in \mathbb{R}^m, n < m\}$.

Gesucht wird eine äquivalente Paarweise orthogonale Menge

$\{\vec{q}_i | i=1, \dots, n, \vec{q}_i \in \mathbb{R}^m, n < m\}$.

Das "Modifizierte Gram-Schmidt-Orthogonalisierungsverfahren", kurz M.G.S.-Verfahren, mit $\vec{a}_i^1 = \vec{a}_i$ ($i=1, \dots, n$) lautet:

$$k = 1, 2, \dots, n$$

$$\vec{q}_k = \vec{a}_k^k$$

$$d_k = \|\vec{q}_k\|_2^2$$

$$\left. \begin{aligned} r_{kj} &= \langle \vec{q}_k, \vec{a}_j^k \rangle / d_k \\ \vec{a}_j^{k+1} &= \vec{a}_j^k - r_{kj} \vec{q}_k \end{aligned} \right\} k+1 \leq j \leq n$$

Wird das M.G.S.-Verfahren auf die Spaltenvektoren einer nicht singulären $(m \times n)$ -Matrix A ($\text{Rang}(A)=n$) angewendet, so ergeben sich eine spaltenorthogonale $(m \times n)$ -Matrix Q und eine Rechtsdreiecksmatrix R mit allen Diagonalelementen gleich 1 derart, dass für die Matrizen Q und R gilt:

$$QR = A \quad (I.7.1)$$

$$Q^T Q = \begin{bmatrix} d_1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & d_n \end{bmatrix} =: D \quad (I.7.2)$$

I.8. Definition.

Es sei A eine $(m \times n)$ -Matrix. Die Gleichungen

$$AXA = A \quad (1.8.1)$$

$$XAX = X \quad (1.8.2)$$

$$(AX)^T = AX \quad (1.8.3)$$

$$(XA)^T = XA \quad (1.8.4)$$

besitzen genau eine Lösung X . Diese Lösung heisst die Pseudoinverse von A , bezeichnet als $X=A^+$ [14,15].

I.9. Darstellung der Pseudoinversen A^+ .

a) Ist A quadratisch nicht singulär, so gilt

$$A^+ = A^{-1}. \quad (1.9.1)$$

b) Ist A eine $(m \times n)$ -Matrix mit $\text{Rang}(A)=n$ und $m > n$, so kann A^+ mit Hilfe der Matrizen U, S, V (siehe I.5.1) folgendermassen dargestellt werden:

$$A^+ = VS^+U^T; \quad (1.9.2)$$

denn es gilt:

$$AA^+A = USV^T VS^+U^T USV^T = USV^T = A$$

$$A^+AA^+ = VS^+U^T USV^T VS^+U^T = VS^+U^T = A^+$$

$$AA^+ = USV^T VS^+U^T = UU^T = (UU^T)^T = (AA^+)^T$$

$$A^+A = VS^+U^T USV^T = VV^T = (VV^T)^T = (A^+A)^T.$$

c) Ist A eine $(m \times n)$ -Matrix mit $\text{Rang}(A)=n$ und $m > n$, so kann A^+ mit Hilfe der Matrizen Q, R, D (siehe I.7.1-I.7.2) wie folgt dargestellt werden:

$$A^+ = R^{-1} D^{-1} Q^T. \quad (1.9.3)$$

In der Tat gelten die Eigenschaften (I.8.1-I.8.4):

$$AA^{\dagger}A = QRR^{-1}D^{-1}Q^TQR = QR = A$$

$$A^{\dagger}AA^{\dagger} = R^{-1}D^{-1}Q^TQRR^{-1}D^{-1}Q^T = A^{\dagger}$$

$$AA^{\dagger} = QRR^{-1}D^{-1}Q^T = QD^{-1}Q^T = (QD^{-1}Q^T)^T = (AA^{\dagger})^T$$

$$A^{\dagger}A = R^{-1}D^{-1}Q^TQR = R^{-1}R = (A^{\dagger}A)^T.$$

d) Es sei A eine (m x n)-Matrix mit Rang(A) = n < m, $P = \begin{bmatrix} A \\ I_n \end{bmatrix}$

die erweiterte Matrix von A, $\vec{p}_i$ (i=1,...,n) die Spaltenvektoren der Matrix P und $\vec{a}_i$ (i=1,...,n) die Spaltenvektoren von A. Auf die Matrix P wird das M.G.S.-Verfahren angewendet, und die Operationen werden auf die ganze ((m+n) x n)-Matrix erweitert. Der Prozess wird wie folgt beschrieben:

$$k = 1, 2, \dots, n$$

$$\left. \begin{array}{l} \vec{p}_i^1 = \vec{p}_i \\ \vec{a}_i^1 = \vec{a}_i \end{array} \right\} \quad i = 1, \dots, n$$

$$\vec{q}_k^1 = \vec{p}_k^1$$

$$d_k = \|\vec{a}_k^1\|_2^2$$

$$\left. \begin{array}{l} r_{kj} = \langle \vec{a}_k^1, \vec{a}_j^1 \rangle / d_k \\ \vec{p}_j^{k+1} = \vec{p}_j^1 - r_{kj} \vec{q}_k^1 \end{array} \right\} \quad k+1 \leq j \leq n$$

Das Verfahren, angewendet auf die Matrix P, liefert die Matrix

$$Q' = \begin{bmatrix} Q \\ I \end{bmatrix}$$

und es gilt:

$$Q'^T Q' = D = \begin{bmatrix} d_1, & \dots, & 0 \\ \vdots & & \vdots \\ 0, & \dots, & d_n \end{bmatrix} \quad (I.9.4)$$

$$A^T = Q \quad (I.9.5)$$

$$A^{\dagger} = TD^{-1}Q^T \quad (I.9.6)$$

[2, 25].

I.10. Bemerkungen.

a) Im Gegensatz zu [27] werden bei (I.9.c) und (I.9.d) die Vektoren $\vec{q}_i$ ($i=1, \dots, n$) der Matrix Q nicht normiert. Die Vorteile sind:

— erstens kann mit rationalen Operationen gearbeitet werden, denn für den k-ten Schritt gilt:

$$\vec{a}_j^k - r'_{kj} \vec{q}_k' = \vec{a}_j^k - r_{kj} \vec{q}_k$$

(mit $r'_{kj} = d_k^{\frac{1}{2}} r_{kj}$, $\vec{q}_k' = d_k^{-\frac{1}{2}} \vec{q}_k$),

— zweitens kann auf n Quadratwurzeln und mxn Divisionen verzichtet werden.

b) Aus (I.9.3) folgt:

$$RA^{\dagger} = D^{-1} Q^T, \tag{I.10.1}$$

Aus (I.10.1) wird ersichtlich, dass die Matrix $A^{\dagger}$ durch Auflösung des linearen Gleichungssystems (I.10.1) durch Rückwärtseinsetzen bestimmt werden kann [27]. Der Vergleich (I.10.1) mit (I.9.6) zeigt, dass bei (I.10.1) die Berechnung der Matrix T aus I_n vermieden wird.

c) Es gibt eine Reihe von Algorithmen, die die Pseudoinverse einer Matrix liefern. Die Mehrheit dieser Algorithmen basiert auf den Methoden der Zerlegung einer Matrix in andere einfachere Matrizen. Die Arbeit von Wilkinson und Peters [31] stellt eine systematische Darstellung solcher Methoden dar, und in der Arbeit von Shinozaki [28] sind die verschiedenen direkten Methoden zur Berechnung pseudoinverser Matrizen klassifiziert und getestet worden.

I.11. Rundungsfehler bei Gleitpunktrechnung.

Es werden die Rundungsfehler einfacher Gleitpunktoperationen angegeben, wie sie zur Fehlerdiskussion des Rechenprozesses benötigt werden.

Auf einem Digitalrechner mit einer Mantissenlänge von t Dualstellen ist die Abschätzung für den relativen Rundungsfehler ϵ einer reellen Zahl a

$$|\epsilon| \leq 2^{-t} \quad . \quad (I.11.1)$$

Dabei liegt ϵ gewöhnlich im Bereich $[10^{-16}, 10^{-6}]$.

I.12. Zusammengesetzte Gleitpunktoperationen.

Sind a und b zwei normalisierte Gleitpunktzahlen, dann bezeichnen $gl(a+b)$, $gl(a-b)$, $gl(ab)$, $gl(a/b)$ diejenigen normalisierten Gleitpunktzahlen, die sich aus den entsprechenden Gleitpunktoperationen ergeben.

Es ist:

$$gl(a+b) = (a+b)(1+\epsilon_1) \quad (I.12.1)$$

$$gl(a-b) = (a-b)(1-\epsilon_2) \quad (I.12.2)$$

$$gl(ab) = (ab)(1+\epsilon_3) \quad (I.12.3)$$

$$gl(a/b) = (a/b)(1+\epsilon_4) \quad (I.12.4)$$

mit

$$|\epsilon_i| \leq 2^{-t} \quad (\text{binär}) \quad (i=1, 2, 3, 4) \quad (I.12.5)$$

$$|\epsilon_i| \leq \frac{1}{2} 10^{1-t} \quad (\text{dezimal}) \quad (I.12.6)$$

(Assoziativ-und Distributivgesetz gelten nicht.)

Für den Fehler, der bei der Berechnung eines mehrfachen Produktes und eines inneren Produktes mit Gleitpunktoperationen auftritt, gelten nach Wilkinson [32] folgende Abschätzungen:

I.13.

Für den relativen Fehler ϵ des mehrfachen Produktes

$$P_n = gl(x_1 \cdot x_2 \dots x_n) = (x_1 \cdot x_2 \dots x_n)(1 + \epsilon) \quad (I.13.1)$$

gilt:

$$|\epsilon| \leq \sqrt{n} 2^{-t} \quad (I.13.2)$$

I.14.

Für den Fehler, der bei der Berechnung eines inneren Produktes

$$S_n = gl(\langle \vec{x}, \vec{y} \rangle) = gl(x_1 y_1 + \dots + x_n y_n) \quad (I.14.1)$$

auftritt, gilt:

$$s_n = x_1 y_1 (1 + \rho_1) + \dots + x_n y_n (1 + \rho_n) \quad (I.14.2)$$

mit

$$(1 - 2^{-t})^{n-r+2} \leq 1 + \rho_r \leq (1 + 2^{-t})^{n-r+2} \quad r=1, \dots, n \quad (I.14.3)$$

Wenn $2^{-t} r < 0.1$ gilt, dann können Ungleichungen der Form

$$(1 - 2^{-t})^r \leq 1 + \rho \leq (1 + 2^{-t})^r \quad (I.14.4)$$

in der Form

$$|\rho| \leq r 2^{-t_1} \quad (I.14.5)$$

geschrieben werden, wobei

$$t_1 = t - 0.08406 \quad (I.14.6)$$

ist.

II. UEBERSICHT UEBER DAS TSCHEBYSCHEFF- AUSGLEICHSPROBLEM.

II.1. Problemstellung.

Es sei ein lineares reelles Gleichungssystem vorgelegt:

$$r_i(\vec{x}) = b_i - \sum_{k=1}^n a_{ik} x_k \quad i=1,2,\dots,m > n. \quad (\text{II.1.1})$$

Das T.A.P. besteht darin, $\vec{x}^T = (x_1, \dots, x_n)$ so zu bestimmen, dass

$$\max_{1 \leq i \leq m} |r_i(\vec{x})|$$

möglichst klein wird.

Hierfür kann man auch

$$\|\vec{b} - A\vec{x}\|_{\infty} = \text{Min}, \quad (\text{II.1.2})$$

schreiben, wobei A eine reelle (m x n)-Matrix und $\vec{b} \in \mathbb{R}^m$, $\vec{x} \in \mathbb{R}^n$ ist.

Folgende Voraussetzungen sollen erfüllt sein:

- a) Die Spaltenvektoren $\vec{a}^i$ ($i=1, \dots, n$) der Matrix A sind linear unabhängig.
- b) $\vec{b}$ ist nicht Linearkombination von $\vec{a}^1, \dots, \vec{a}^n$.
- c) Jede n-reihige quadratische Teilmatrix von A ist nicht singulär (Haarsche-Bedingung).

Die Fragen

der Existenz
der Eindeutigkeit
der Charakterisierung

von Minimallösungen für das Problem (II.1.2) werden in [6,7,8,9,19, 23] abschliessend behandelt.

Ein Punkt $\vec{x}_T$, für den die Bedingung (II.1.2) gilt, nennt man Tschebyscheff-Punkt, kurz T-Punkt. Allgemein wird auch hier der Buchstabe T bei ähnlichen Ausdrücken als Abkürzung für "Tschebyscheff" verwendet.

II.2. Theoretische Grundlagen.

- Es sind $\vec{a}_i^T = (a_{i1}, \dots, a_{in})$ ($i=1, \dots, m$) die Zeilenvektoren der $(m \times n)$ -Matrix A.

- I_0 ist die Indexmenge $I_0 = \{1, 2, \dots, m\}$.

- I eine Menge von $n+1$ verschiedenen Indizes aus I_0
 $I = \{i_1, \dots, i_{n+1}\}$.

- $o(I)$ bedeutet die Anzahl der Elemente einer Menge I.

- Als Referenz [I] wird eine Auswahl von $n+1$ unter den gegebenen m Gleichungen des Systems (II.1.1) bezeichnet mit Zeilenindizes $i_k \in I$.

- Die der Referenz [I] entsprechende Teilmatrix von A wird mit A_I bezeichnet.

- Mit $\vec{b}_I$, $\vec{f}_I$ werden die Teilvektoren bezeichnet, die der Referenz [I] zugeordnet sind.

- Ist die Haarsche-Bedingung erfüllt, so existiert ein Vektor $\vec{f}_I^T = (f_{i_1}, \dots, f_{i_{n+1}})$ mit allen Komponenten $f_{i_k} \neq 0$ $i_k \in I$, der die Gleichung

$$\vec{f}_I^T A_I = 0 \quad (\text{II.2.1})$$

erfüllt.

- Ein Vektor $\vec{x}$ wird Referenzvektor genannt, wenn entweder

$$\text{sgn}(\vec{f}_I) = \text{sgn}(\vec{f}_I^T) \quad (\text{II.2.2})$$

oder

$$\text{sgn}(\vec{f}_I) = -\text{sgn}(\vec{f}_I^T) \quad (\text{II.2.3})$$

mit

$$\vec{b}_I - A_I \vec{x} = \vec{f}_I$$

gilt.

- Es sei $\vec{g}_I = \text{sgn}(\vec{f}_I^T)$, dann ist die $(n+1) \times (n+1)$ -Matrix

$|A_I| \vec{g}_I$ nichtsingulär, und es existiert genau eine Lösung

$$\left[\begin{array}{c} \vec{x}_{I_T} \\ h_I \end{array} \right] \text{ des Gleichungssystems } \vec{b}_I - A_I \vec{x} = h_I \vec{g}_I. \quad (\text{II.2.4})$$

- Der Vektor $\vec{x}_{I_T}$ wird als T-Lösung oder Zentrum der Referenz [I] bezeichnet; der Betrag der Grösse h_I heisst Referenzabweichung der Referenz [I] und wird durch die Formel

$$h_I = \frac{\langle \vec{I}_I, \vec{b}_I \rangle}{\| \vec{I}_I \|_1} \quad (\text{II.2.5})$$

gegeben [20,29].

- Die T-Lösung des Systems (II.1.1) ist auch T-Lösung irgendeiner Referenz des Systems (II.1.1) [6].

II.3. Austauschsatz.

Es sei eine Referenz [I] vorgelegt, ein zugehöriger Referenzvektor $\vec{x}$, sowie der Vektor $\vec{a}_k^T$ mit $k \in I_0$.

Dann kann aus der Referenz [I] und aus dem Index k eine neue Referenz [I'] ($I' = I \setminus \{k\} / \{i_j\}$, $i_j \in I$) gebildet werden, für die $\vec{x}$ auch Referenzvektor ist [6,29].

- Mit Hilfe des Austauschsatzes kann nun das T.A.P. gelöst werden.

Es wird mit irgend einer Referenz [I] gestartet, und es werden ihre T-Lösung $\vec{x}_{I_T}$, sowie ihre Referenzabweichung

h_I berechnet. Dann werden alle Residuen $r_i(\vec{x}_{I_T})$, $i \in I_0$ ermittelt. Gilt

$$|r_i(\vec{x}_{I_T})| \leq |h_I| \quad \text{für } \forall i \in I_0,$$

so wird die Konstruktion abgebrochen; denn in diesem Fall ist $\vec{x}_{I_T}$ die T-Lösung des Systems (II.1.1). Andernfalls gibt es mindestens eine Gleichung k , so dass

$$|r_k(\vec{x}_{I_T})| > |h_I| \quad k \in I_0 / I$$

ist.

Nach dem Austauschsatz existiert eine neue (zweite) Referenz

$[I']$, für welche $\vec{x}_I$ auch noch Referenzvektor ist. Es ist nun wesentlich, dass für die Referenzabweichung $h_{I'}$ der neuen Referenz $[I']$ die Beziehung

$$|h_{I'}| > |h_I|$$

stets gilt, das heisst, die Referenzabweichung wird bei jedem Austausch sicher vergrössert.

II.4. Ausgleichung nach kleinsten Quadraten.

Es seien I , A_I , $\vec{b}_I$ wie in (II.2) definiert.

a) Das Minimum des Funktionals

$$|\vec{b}_I - A_I \vec{x}_I|^2$$

wird für

$$\vec{x}_{IQ} = A_I^\dagger \vec{b}_I$$

angenommen $[14, 15, 21, 31]$.

b) Der Residuenvektor

$$\vec{r}_I(\vec{x}_{IQ}) = \vec{b}_I - A_I \vec{x}_{IQ} \quad (\text{II.4.1})$$

genügt der Gleichung

$$\vec{r}_I(\vec{x}_{IQ})^\dagger A_I = 0 \quad (\text{II.4.2})$$

Es ist:

$$(\vec{b}_I - A_I A_I^\dagger \vec{b}_I)^\dagger A_I = \vec{b}_I^\dagger A_I - \vec{b}_I^\dagger (A_I A_I^\dagger)^\dagger A_I = 0$$

c) Es gilt:

$$B_I \vec{b}_I = \vec{r}_I(\vec{x}_{IQ}) \quad (\text{II.4.3})$$

$$B_I' = B_I, \quad (\text{II.4.4})$$

wobei

$$B_I = I_{n+1} - A_I A_I^\dagger \text{ ist.}$$

II.5. Zusammenhang zwischen den Lösungen $\vec{x}_{IQ}$ und $\vec{x}_{IT}$

Zwischen der Lösung $\vec{x}_{IQ}$ des Problems

$$\min_{\vec{x} \in \mathbb{R}^n} \|\vec{b}_I - A_I \vec{x}\|_2 \quad (II.5.1)$$

und der l -Lösung $\vec{x}_{IT}$ des Problems

$$\min_{\vec{x} \in \mathbb{R}^n} \|\vec{b}_I - A_I \vec{x}\|_\infty \quad (II.5.2)$$

besteht die Beziehung

$$\vec{x}_{IT} = \vec{x}_{IQ} - h_I A_I^+ \vec{g}_I \quad (II.5.3)$$

[6,10,9].

Dabei ist wegen (II.4.2)

$$\vec{g}_I = \text{sgn}(\vec{b}_I - A \vec{x}_{IQ}) = \text{sgn}(\vec{r}_I(\vec{x}_{IQ})) = \text{sgn}(\vec{r}_{IQ}) = \text{sgn}(\vec{r}_I)$$

und

$$h_I = \frac{\langle \vec{b}_I, \vec{r}_{IQ} \rangle}{\|\vec{r}_{IQ}\|_1} = \frac{\langle \vec{r}_I, \vec{r}_{IQ} \rangle}{\|\vec{r}_{IQ}\|_1}, \quad (II.5.4)$$

weil

$$\langle \vec{b}_I, \vec{r}_{IQ} \rangle = \langle \vec{r}_I, \vec{b}_I - A \vec{x}_{IQ} \rangle = \langle \vec{r}_I, \vec{r}_{IQ} \rangle$$

gilt.

II.6. Das Austauschkriterium für die Vergrößerung der Referenzabweichung.

Durch das Kriterium

$$\frac{y_j}{f_j} = \max_{i \in I} \frac{y_i}{f_i}, \quad (II.6.1)$$

wobei $\vec{y}^T = (y_{i_1}, \dots, y_{i_{n+1}})$ eine Lösung des Systems

$$\dot{y}^T A_I = \operatorname{sgn}(r_k(\vec{y}_{I_I})) \cdot \vec{a}_k^T \quad \forall k \in I_0 \quad (11.6.2)$$

ist, wird die i_j -te Gleichung bestimmt, welche von der Referenz [I] ausscheidet und zugleich die Referenzabweichung $h_{I'}$ der Referenz [I'] sicher vergrößert [10].

III. AENDERUNGSFORMEL DER PSEUDOINVERSEN MATRIX A_I^+ . LOESUNGSFORMELN.

In diesem Kapitel werden Relationen zwischen dem Vektor

$$\vec{f}_I = \frac{\vec{r}_{IQ}}{\|\vec{r}_{IQ}\|_2} \quad \text{und den Spaltenvektoren der Matrix } B_I = I_{n+1} - A_I A_I^+$$

hergestellt. Weiter wird eine Aenderungsformel für die Bestimmung der Matrix A_I^+ hergeleitet und eine Formel, die die Lösung einer neuen Referenz nach der Methode der kleinsten Quadrate angibt, wenn die Lösung der alten Referenz bekannt ist.

III.1. Satz.

Es sei $\text{Rang}(A_I) = n$, $\vec{r}_{IQ} := \vec{r}_I(\vec{x}_{IQ}) \neq 0$ und $\vec{f}_I = \frac{\vec{r}_{IQ}}{\|\vec{r}_{IQ}\|_2}$.

Dann gelten:

$$a) \quad \vec{f}_I^T A_I = 0 \quad (\text{III.1.1})$$

$$b) \quad B_I = \vec{f}_I \vec{f}_I^T \quad (\text{III.1.2})$$

$$c) \quad \text{sp}(A_I A_I^+) = n \quad (\text{III.1.3})$$

$$d) \quad \vec{r}_{IQ} = \vec{f}_I (h_I \|\vec{f}_I\|_1) \quad (\text{III.1.4})$$

$$e) \quad \|\vec{f}_I\|_1 \leq \sqrt{n+1} \quad (\text{III.1.5})$$

Beweis:

Zu a). Aus (II.4.2) und $\vec{f}_I = \frac{\vec{r}_{IQ}}{\|\vec{r}_{IQ}\|_2}$ folgt (III.1.1).

Zu b). Die erweiterte Matrix $A_I' = \left[\begin{array}{c|c} \vec{1}_I & A_I \end{array} \right]$ ist nicht
singulär, weil $\text{Rang}(A_I) = n$ und $\vec{r}_{I_Q} \neq 0$ ist.
Es sei

$$A_I'^{-1} = \left[\begin{array}{c} \vec{u}^T \\ B_0 \end{array} \right] \quad (III.1.6)$$

Dann folgen aus

$$A_I'^{-1} A_I' = \left[\begin{array}{cc} \vec{u}^T \vec{1}_I & \vec{u}^T A_I \\ B_0 \vec{1}_I & B_0 A_I \end{array} \right] = I_{n+1} \quad (III.1.7)$$

die Gleichungen

$$\vec{u}^T \vec{1}_I = 1 \quad (III.1.8)$$

$$\vec{u}^T A_I = 0 \quad (III.1.9)$$

$$B_0 \vec{1}_I = 0 \quad (III.1.10)$$

$$B_0 A_I = I_n \quad (III.1.11)$$

Folglich ist

$$\vec{u} = \vec{1}_I \quad (III.1.12)$$

$$B_0 = A_I^{-1} \quad (III.1.13)$$

Aus

$$A_I' A_I'^{-1} = I_{n+1} \quad (\text{III.1.14})$$

folgt, dass

$$A_I A_I^\dagger + \vec{f}_I \vec{f}_I^\dagger = I_{n+1} \quad (\text{III.1.15})$$

ist; d.h.

$$B_I = I_{n+1} - A_I A_I^\dagger = \vec{f}_I \vec{f}_I^\dagger \quad (\text{III.1.16})$$

oder

$$f_j^2 = 1 - \langle \vec{a}_j, \vec{a}^{j\dagger} \rangle \quad j \in I \quad (\text{III.1.17})$$

$$f_i f_j = -\langle \vec{a}_i, \vec{a}^{j\dagger} \rangle = -\langle \vec{a}_j, \vec{a}^{i\dagger} \rangle \quad i, j \in I \text{ und } i \neq j, \quad (\text{III.1.18})$$

wobei $\vec{a}^{j\dagger}$ die Spaltenvektoren der Matrix $A_I^\dagger$ sind.

Zu c). Weil die Spur einer Matrix die Eigenschaft der Linearität hat, folgt

$$\begin{aligned} \text{sp}(A_I A_I^\dagger + \vec{f}_I \vec{f}_I^\dagger) &= \text{sp}(A_I A_I^\dagger) + \text{sp}(\vec{f}_I \vec{f}_I^\dagger) = \\ &= \text{sp}(A_I A_I^\dagger) + 1 = n+1. \end{aligned}$$

Zu d). Aus (II.4.3), (III.1.2) und (II.2.5) folgt:

$$\vec{f}_{I_Q} = B_I \vec{b}_I = \vec{f}_I \vec{f}_I^\dagger \vec{b}_I = \vec{f}_I (\langle \vec{b}_I | | \vec{f}_I | |_1) \quad (\text{III.1.19})$$

Zu e). Aus (III.1.19) folgt:

$$\langle \vec{f}_{I_Q}, \vec{f}_{I_Q} \rangle = h_I^2 \| \vec{f}_I \|_1^2 \leq \langle \vec{f}_{I_T}, \vec{f}_{I_T} \rangle = h_I^2 (n+1),$$

$$\| \vec{f}_I \|_1 \leq \sqrt{n+1},$$

wobei $\vec{f}_{I_T} = \vec{b}_I - A_I \vec{x}_{I_T}$ ist. $\square$

- Für die Bestimmung des Vektors $\vec{f}_I$ löst Cheney [6] in seinem Algorithmus das Gleichungssystem

$$\vec{f}_I^T A_I = 0$$

und Stiefel [29] das System

$$\vec{f}_I^T L = 0 ,$$

wobei L eine Dreiecksmatrix ist. Die Formeln (III.1.17), (III.1.18) ermöglichen den Vektor $\vec{f}_I$ leicht zu bestimmen, wenn die Matrix A_I^+ bekannt ist.

- Im folgenden wird untersucht, wie bei der Änderung einer Zeile einer Referenzmatrix A_I die Pseudoinverse A_I^+ sich ändert. Gegeben seien die erste Referenz $[I]$

$$I = \{i_1, i_2, \dots, i_j := j, \dots, i_{n+1}\}$$

und die zweite Referenz $[I']$

$$I' = \{i_1, i_2, \dots, i_{j-1}, i_k := k, i_{j+1}, \dots, i_{n+1}\}$$

mit $k \in I_0/I$.

Mit $\vec{a}_i^T$, bzw. $\vec{a}_{i'}^T$ werden die Zeilen der ersten Referenzmatrix A_I , bzw. der zweiten Referenzmatrix $A_{I'}$ und mit $\vec{a}^{i+}$, bzw. $\vec{a}^{i'+}$ die Spalten der Matrix A_I^+ , bzw. $A_{I'}^+$ bezeichnet.

III.2. Satz.

Für die Spaltenvektoren $\vec{a}^{i+}$ der Matrix A_I^+ gilt:

$$\vec{a}^{i'} = \vec{a}^{i+} - y_i \vec{z} - f_i \vec{z}' \quad i \in I / \{j\} \quad j \in I, \quad i' \in I' \quad (\text{III.2.1})$$

$$\vec{a}^{j+} = \vec{z} \quad I' \ni j' := k \quad k \in I_0 / I, \quad (\text{III.2.2})$$

wobei

$$\vec{z} := \frac{1}{\mu \nu} \left[f_j^2 \vec{d} + y_j \vec{a}^{j+} \right] \quad (\text{III.2.3})$$

$$\vec{z}' := \frac{f_j}{\mu \nu} \left[\nu \vec{a}^{j+} - y_j \vec{d} \right] \quad (\text{III.2.4})$$

$$\vec{y}^T = \text{sgn}(r_k(\vec{x}_{I_T})) \vec{a}_k^T A_I^+ \quad (\text{III.2.5})$$

$$\vec{d} = A_I^+ \vec{y} \quad (\text{III.2.6})$$

$$\nu := 1 + \|\vec{y}\|_2^2 \quad (\text{III.2.7})$$

$$\mu := f_j^2 + \frac{y_j^2}{\nu} \quad (\text{III.2.8})$$

ist.

Als Vorbereitung zum Beweis dieses Satzes wird folgendes Lemma vorangestellt.

III.3. Lemma.

Es sei $\vec{x}_{I_T}$ T-Lösung der ersten Referenz [I] und $\text{sgn}(r_k(\vec{x}_{I_T})) = +1$.

Für die Komponenten y_i , $i \in I$ des Vektors $\vec{y}^T = \vec{a}_k^T A_I^+$ gilt

$$a) \quad y_i = \langle \vec{a}_i, \vec{d} \rangle \quad \forall i \in I \quad (\text{III.3.1})$$

$$b) \quad \|\vec{y}\|_2^2 = \langle \vec{a}_k, \vec{d} \rangle, \quad (\text{III.3.2})$$

wobei

$$\vec{d} = A_I^+ \vec{y}$$

ist.

Beweis:

Zu a). Aus

$$\vec{y}^T = \vec{a}_k^T A_I^\dagger$$

und

$$A_I^\dagger = (A_I^\dagger A_I^\dagger)^T A_I^T$$

folgt:

$$y_i = \langle \vec{a}_k, A_I^\dagger A_I^\dagger \vec{a}_i \rangle = \langle \vec{a}_i, A_I^\dagger \vec{y} \rangle = \langle \vec{a}_i, \vec{d} \rangle \quad i \in I.$$

Zu b). Es ist

$$\|\vec{y}\|_2^2 = \langle \vec{y}, \vec{y} \rangle = \vec{a}_k^T A_I^\dagger \vec{y} = \langle \vec{a}_k, \vec{d} \rangle. \quad \square$$

Beweis des Satzes III.2.

Für die Matrizen $A_{I'}^\dagger$, $C_{I'} := A_{I'}^T A_{I'}$ gelten:

$$A_{I'}^\dagger = A_{I'} C_{I'}^{-1} \quad (\text{III.2.9})$$

$$\begin{aligned} C_{I'} &= A_I^T A_I + \vec{a}_k \vec{a}_k^T - \vec{a}_j \vec{a}_j^T = \\ &= C_I + \vec{a}_k \vec{a}_k^T - \vec{a}_j \vec{a}_j^T = T_I - \vec{a}_j \vec{a}_j^T, \end{aligned} \quad (\text{III.2.1})$$

wobei

$$T_I := C_I + \vec{a}_k \vec{a}_k^T$$

ist.

Aus (I.6.1) und (I.6.2) folgt:

$$T_I^{-1} = C_I^{-1} - \frac{C_I^{-1} \vec{a}_k \vec{a}_k^T C_I^{-1}}{1 + \vec{a}_k^T C_I^{-1} \vec{a}_k} \quad (\text{III.2.1})$$

$$C_{I'}^{-1} = T_I^{-1} + \frac{T_I^{-1} \vec{a}_j \vec{a}_j^T T_I^{-1}}{1 - \vec{a}_j^T T_I^{-1} \vec{a}_j}. \quad (\text{III.2.12})$$

Aus Lemma III.3. folgt:

$$\begin{aligned} 1 + \vec{a}_k^T C_I^{-1} \vec{a}_k &= 1 + \vec{a}_k^T A_I^T A_I^T \vec{a}_k = 1 + \langle \vec{a}_k, \vec{d} \rangle = \\ &= 1 + \|\vec{y}\|_2^2 =: v \end{aligned} \quad (\text{III.2.13})$$

und

$$\begin{aligned} 1 - \vec{a}_j^T C_I^{-1} \vec{a}_j &= 1 - \vec{a}_j^T C_I^{-1} \vec{a}_j + \frac{\vec{a}_j^T C_I^{-1} \vec{a}_k \vec{a}_k^T C_I^{-1} \vec{a}_j}{1 + \vec{a}_k^T C_I^{-1} \vec{a}_k} = \\ &= 1 - \langle \vec{a}_j, \vec{a}^{j+} \rangle + \frac{1}{v} \vec{a}_j^T \vec{d} \vec{d}^T \vec{a}_j . \end{aligned}$$

Aus (III.1.17) und Lemma III.3 folgt:

$$1 - \vec{a}_j^T C_I^{-1} \vec{a}_j = f_j^2 + \frac{y_j^2}{v} =: \mu \quad . \quad (\text{III.2.14})$$

Weiter gelten:

$$C_I^{-1} A_I^T = (\vec{a}^{i_1+}, \dots, \vec{a}^{i_{j-1}+}, \vec{d}, \vec{a}^{i_{j+1}+}, \dots, \vec{a}^{i_{n+1}+}) =: G \quad (\text{III.2.15})$$

$$A_I \vec{d} = A_I A_I^T \vec{y} = \begin{pmatrix} y_{i_1} \\ \vdots \\ y_{i_{j-1}} \\ \langle \vec{y}, \vec{y} \rangle \\ y_{i_{j+1}} \\ \vdots \\ y_{i_{n+1}} \end{pmatrix} =: \vec{s} \quad . \quad (\text{III.2.16})$$

Aus (III.2.11), (III.2.15), (III.2.16) folgt:

$$A_I A_I^T = G^T - \frac{1}{v} \vec{s} \vec{s}^T \quad . \quad (\text{III.2.17})$$

Weiter gilt:

$$\vec{a}_J^T I_I^{-1} = \vec{a}_J^T C_I^{-1} = \frac{1}{\mu} \vec{a}_J^T \vec{d} \vec{d}^T = (\vec{a}^{j+})^T = \frac{1}{\mu} y_J \vec{d}^T, \quad (\text{III.2.18})$$

und daraus folgt:

$$\vec{a}_J^T I_I^{-1} A_{I'}^T = (\vec{a}^{j+})^T A_{I'}^T = \frac{1}{\mu} y_J \vec{s}^T. \quad (\text{III.2.19})$$

Aus (III.2.11), (III.2.12), (III.2.14) und (III.2.17- III.2.19) folgt:

$$\begin{aligned} A_{I'}^{\dagger T} &= A_{I'} C_{I'}^{-1} = A_{I'}^T I_I^{-1} + \frac{1}{\mu} A_{I'}^T I_I^{-1} \vec{a}_J \vec{a}_J^T I_I^{-1} = \\ &= G^T - \frac{1}{\nu} \vec{s} \vec{d}^T + \frac{1}{\mu} \left[A_{I'}^T \vec{a}^{j+} - \frac{y_J}{\nu} \vec{s} \right] \left[\vec{a}^{j+} - \frac{y_J}{\nu} \vec{d} \right]^T = \\ &= G^T + \frac{1}{\nu} \vec{s} \vec{d}^T \left(\frac{y_J^2}{\mu \nu} - 1 \right) - \frac{y_J}{\mu \nu} \vec{s} \vec{a}^{j+T} + \frac{1}{\mu} A_{I'}^T \vec{a}^{j+} \vec{a}^{j+T} - \\ &\quad - \frac{y_J}{\mu \nu} A_{I'}^T \vec{a}^{j+} \vec{d}^T = \\ &= G^T - \frac{f^2}{\mu \nu} \vec{s} \vec{d}^T - \frac{y_J}{\mu \nu} \vec{s} \vec{a}^{j+T} + \frac{1}{\mu} A_{I'}^T \vec{a}^{j+} \vec{a}^{j+T} - \frac{y_J}{\mu \nu} A_{I'}^T \vec{a}^{j+} \vec{d}^T = \\ &= G^T - \frac{s}{\mu \nu} \left[f_j^2 \vec{d} + y_J \vec{a}^{j+} \right]^T + \frac{A_{I'}^T \vec{a}^{j+}}{\mu \nu} \left[\nu \vec{a}^{j+} - y_J \vec{d} \right]^T. \quad (\text{III.2.20}) \end{aligned}$$

Aus (III.1.18) und Lemma III.3 folgt:

$$A_{I'}^{\dagger T} \vec{a}^{j+} = -f_j \begin{vmatrix} f_{i_1} \\ \vdots \\ f_{i_{j-1}} \\ -\frac{y_J}{f_j} \\ f_{i_{j+1}} \\ \vdots \\ f_{i_{n+1}} \end{vmatrix} =: -f_j \vec{t}_o. \quad (\text{III.2.21})$$

Wird (III.2.21), (III.2.3), (III.2.4) in (III.2.20) eingesetzt, so gilt:

$$A_{I'}^{\dagger T} = G^T - \vec{s} \vec{z} - \vec{t}_o \vec{z}^T. \quad (\text{III.2.22})$$

Weil

$$\begin{aligned}
 \vec{a} - \frac{\langle \vec{y}, \vec{y} \rangle}{f_j} \vec{z} + \frac{y_j}{f_j} \vec{z}' &= \vec{a} - \frac{\langle \vec{y}, \vec{y} \rangle}{\mu\nu} \left[f_j^2 \vec{a} + y_j \vec{a}^{j+} \right] + \\
 &+ \frac{y_j}{\mu\nu} \left[y_j \vec{a}^{j+} - y_j \vec{a} \right] = \\
 &= \left(1 - \frac{\langle \vec{y}, \vec{y} \rangle}{\mu\nu} f_j^2 - \frac{y_j^2}{\mu\nu} \right) \vec{a} + \\
 &+ \left(\frac{y_j}{\mu} - \frac{\langle \vec{y}, \vec{y} \rangle}{\mu\nu} \frac{y_j}{\mu\nu} \right) \vec{a}^{j+} = \\
 &= \frac{f_j^2}{\mu\nu} \vec{a} + \frac{y_j}{\mu\nu} \vec{a}^{j+} =: \vec{z}
 \end{aligned}$$

ist, folgt:

$$\vec{a}^{i+} = \vec{a}^{i+} - y_i \vec{z} - f_i \vec{z}' \quad i \in I \setminus \{j\}, j \in I, i \in I' \quad (\text{III.2.23})$$

$$\vec{a}^{j+} = \vec{z} \quad I \ni j' := k \quad k \in I_0/I, \quad (\text{III.2.24})$$

wobei die linke Seite der Gleichungen (III.2.23 - III.2.24) die Spaltenvektoren der Matrix A_I^+ sind. $\square$

III.4. Korollar

Die Matrix A_I^+ kann berechnet werden, wenn $\mu\nu \neq 0$, d.h. $f_j^2 + \frac{y_j^2}{\nu} \neq 0$ ist.

III.5. Satz.

Für die Komponenten $f_i, i \in I'$ des Vektors $\vec{f}_I$ der zweiten Referenz $[I']$ gelten:

$$a) \quad f_{j'} = \frac{f_j^2}{\mu\nu} \quad I \ni j' := k \quad j \in I \quad (\text{III.5.1})$$

$$b) \quad f_{i'} = f_{j'} \left[\frac{y_j}{f_j} f_i - y_i \right] \quad i \in I', I \ni i \neq j \in I \quad (\text{III.5.2})$$

$$c) \quad \|\vec{f}_I\|_2^2 = 1$$

Beweis :

Unter Anwendung von (III.1.17), (III.1.18) und des Satzes III.2. hat man:

Zu a)

$$\begin{aligned} f_{j'}^2 &= 1 - \langle \vec{a}_k, \vec{z} \rangle = 1 - \frac{1}{\mu\nu} f_j^2 \langle \vec{y}, \vec{y} \rangle - \frac{y_j^2}{\mu\nu} \\ &= 1 - \frac{1}{\mu\nu} f_j^2 (\nu - 1) + \frac{f_j^2}{\mu} - 1 = \frac{f_j^2}{\mu\nu} \quad j \in I, j' \in I' \end{aligned}$$

Zu b)

$$\begin{aligned} f_i f_{j'} &= - \langle \vec{a}_i, \vec{z} \rangle = \frac{1}{\mu\nu} \left(y_j f_i f_j - f_j y_i \right) = \\ &= \frac{f_j^2}{\mu\nu} \left(\frac{y_j}{f_j} f_i - y_i \right) = f_{j'}^2 \left(\frac{y_j}{f_j} f_i - y_i \right), \end{aligned}$$

d.h.

$$f_{i'} = f_{j'} \left(\frac{y_j}{f_j} f_i - y_i \right) \quad i \in I \setminus \{j\}, i' \in I', I' \ni j' := k.$$

Zu c) Aus a), b) und $\langle \vec{y}, \vec{1}_{I'} \rangle = 0$ folgt:

$$\begin{aligned} \|\vec{1}_{I'}\|_2^2 &= \frac{f_j^2}{\mu\nu} + \frac{f_j^2}{\mu\nu} \left(\langle \vec{y}, \vec{y} \rangle + \left(\frac{y_j}{f_j} \right)^2 \right) = \\ &= \frac{f_j^2}{\mu\nu} \left(\nu + \frac{y_j^2}{f_j^2} \right) = 1. \quad \square \end{aligned}$$

Die Berechnung des Vektors $\vec{1}_{I'}$ mit Hilfe der Formel (III.5.1), (III.5.2) benötigt n^2 Operationen (1 Oper. = 1 Mult. + 1 Addition) weniger als mit Hilfe der Formel (III.1.7), (III.1.18).

III.6. Satz.

Sei $\vec{x}_{IQ}$ Lösung des Problems

$$\min_{\vec{x} \in \mathbb{R}^n} \|\vec{b}_I - A_I \vec{x}\|_2^2.$$

Dann ist die Lösung $\vec{x}_{IQ}$ des Problems

$$\min_{\vec{x} \in \mathbb{R}^n} \|\vec{b}_I - A_I \vec{x}\|_2^2$$

gegeben durch

$$\vec{x}_{IQ} = \vec{x}_{IQ} + r_k(\vec{x}_{IQ}) \vec{z} - \langle \vec{I}_I, \vec{b}_I \rangle \vec{z}', \quad (\text{III.6.1})$$

wobei

$$\vec{z} = \frac{1}{\mu\nu} \begin{bmatrix} \vec{f}_j \vec{a} + y_j \vec{a}^{j\dagger} \end{bmatrix} \quad (\text{III.6.2})$$

$$\vec{z}' = \frac{\vec{f}_j}{\mu\nu} \begin{bmatrix} \nu \vec{a}^{j\dagger} - y_j \vec{a} \end{bmatrix} \quad (\text{III.6.3})$$

$$r_k(\vec{x}_{IQ}) = b_k - \langle \vec{a}_k, \vec{x}_{IQ} \rangle \quad (\text{III.6.4})$$

sind.

Beweis:

Es ist

$$\vec{x}_{IQ} = A_I^+ \vec{b}_I. \quad (\text{III.6.5})$$

Aus (III.6.2)-(III.6.3) folgt:

$$\vec{a}^{j\dagger} = y_j \vec{z} + \vec{f}_j \vec{z}'. \quad (\text{III.6.6})$$

Und weil

$$r_k(\vec{x}_{IQ}) = b_k - \vec{a}_k^T A_I^+ \vec{b}_I = b_k - \langle \vec{y}, \vec{b}_I \rangle$$

gemäss Lemma III.3. ist, folgt aus (III.2.1-III.2.2) und (III.6.6):

$$\vec{x}_{IQ} = \vec{x}_{IQ} - b_j \vec{a}^{j\dagger} + b_k \vec{z} - \langle \vec{y}, \vec{b}_I \rangle \vec{z} + y_j b_j \vec{z} -$$

$$- \langle \vec{I}_I, \vec{b}_I \rangle \vec{z}' + \vec{f}_j b_j \vec{z}' =$$

$$= \vec{x}_{IQ} + r_k(\vec{x}_{IQ}) \vec{z} - \langle \vec{I}_I, \vec{b}_I \rangle \vec{z}'. \quad \square$$

Im folgenden wird mit h_I^j die Referenzabweichung der neuen Referenz $[I']$ bezeichnet, wobei der obere Index j derjenigen i_j -ten Gleichung entspricht, die aus der Referenz $[I]$ ausscheidet.

III.7. Satz.

Für die Referenzabweichung $h_I^j =: h_{I'}^j$ der zweiten Referenz $[I']$ gilt:

$$h_I^j = \frac{r_k(\vec{x}_{I_Q}) + \frac{y_j}{I_j} \langle \vec{f}_I, \vec{b}_I \rangle}{1 + \sum_{i \in I} |y_i - \frac{y_j}{I_j} f_i|} \quad i, j \in I, \quad i \neq j. \quad (\text{III.7.1})$$

Beweis:

Aus (II.2.5), (III.5.1), (III.5.2) folgt:

$$\begin{aligned} h_I^j &= \frac{\langle \vec{f}_{I'}^j, \vec{b}_{I'}^j \rangle}{\|\vec{f}_{I'}^j\|_1} = \frac{f_{j'}^j b_k - f_{j'}^j \langle \vec{y}, \vec{b}_I \rangle + f_{j'}^j \frac{y_j}{I_j} \langle \vec{f}_I, \vec{b}_I \rangle}{|f_{j'}^j| + |f_{j'}^j| \left(\sum_{i \in I} |y_i - \frac{y_j}{I_j} f_i| \right)} = \\ &= \frac{r_k(\vec{x}_{I_Q}) + \frac{y_j}{I_j} \langle \vec{f}_I, \vec{b}_I \rangle}{1 + \sum_{i \in I} |y_i - \frac{y_j}{I_j} f_i|} \operatorname{sgn}(f_{j'}^j) \quad j' \in I'. \end{aligned}$$

Wird $\operatorname{sgn}(f_{j'}^j) = \operatorname{sgn}(r_k(\vec{x}_{I_Q}) + \frac{y_j}{I_j} \langle \vec{f}_I, \vec{b}_I \rangle)$ festgesetzt (was wegen der Definition des Vektors $\vec{f}_{I'}^j$ stets möglich ist), so wird

$$h_I^j = \frac{r_k(\vec{x}_{I_Q}) + \frac{y_j}{I_j} \langle \vec{f}_I, \vec{b}_I \rangle}{1 + \sum_{i \in I} |y_i - \frac{y_j}{I_j} f_i|} > 0 \quad i, j \in I, \quad i \neq j. \quad \square$$

Der Satz II.6 kann auf Grund der Formel (III.7.1) wie folgt bewiesen werden:

III.8. Satz.

Ist

$$\frac{y_j}{I_j} = \max_{i \in I} \frac{y_i}{I_i} \quad (\text{III.8.1})$$

und

$$r_k(\vec{x}_{I_T}) = \max_{i \in I_0} |r_i(x_{I_T})| > h_I \quad i \in I_0 \quad (\text{III.8.2})$$

dann gilt stets die Ungleichung:

$$h_I^j > \frac{\langle \vec{f}_I, \vec{b}_I \rangle}{\|\vec{f}_I\|_1} = h_I. \quad (\text{III.8.3})$$

Beweis: Indirekter Beweis.

Aus (III.8.1) folgt:

$$\text{sgn}\left(\frac{y_j}{I_j} f_i - y_i\right) = \text{sgn}(f_i) \quad \forall i \in I.$$

Angenommen es gelte

$$h_I^j \leq h_I,$$

dann hat man aus (III.7.1)

$$\frac{r_k(\vec{x}_{I_Q}) + \frac{y_j}{I_j} \langle \vec{f}_I, \vec{b}_I \rangle}{1 + \sum_{i \in I} \left| \frac{y_j}{I_j} f_i - y_i \right|} \leq \frac{\langle \vec{f}_I, \vec{b}_I \rangle}{\|\vec{f}_I\|_1}$$

und daher

$$r_k(\vec{x}_{I_Q}) \|\vec{f}_I\|_1 + \langle \vec{f}_I, \vec{b}_I \rangle \left(\frac{y_j}{I_j} \|\vec{f}_I\|_1 - 1 - \sum_{i \in I} \left| \frac{y_j}{I_j} f_i - y_i \right| + \sum_{i \in I} y_i \text{sgn}(f_i) \right) \leq 0$$

$$r_k(\vec{x}_{I_Q}) + h_I \sum_{i \in I} y_i \text{sgn}(f_i) - h_I \leq 0.$$

Hieraus folgt, dass

$$r_k(\vec{x}_{I_T}) \leq h_I$$

ist. Dies ist ein Widerspruch, weil

$$r_k(\vec{x}_{I_T}) = \max_{i \in I_0} |r_i(\vec{x}_{I_T})| > h_I$$

ist.

Der Satz (III.8) folgt aus diesem Widerspruch. $\square$

In Satz III.1 wurde die Voraussetzung $\vec{r}_{I_Q} := \vec{r}_I(\vec{x}_{I_Q}) \neq 0$ getroffen. Es ist denkbar, dass diese Annahme einmal nicht erfüllt ist.

Im folgenden Satz wird bewiesen, dass diese Annahme höchstens beim Start nicht erfüllt sein kann.

III.9. Satz.

Ist

$$r_k(\vec{x}_{I_T}) > 0 \quad (III.9.1)$$

und

$$r_i(\vec{x}_{I_Q}) = 0 \quad \forall i \in I, \quad (III.9.2)$$

so ist

$$r_i(\vec{x}_{I_Q'}) \neq 0 \quad \forall i \in I'. \quad (III.9.3)$$

Beweis: Indirekter Beweis.

Nimmt man an, dass

$$r_i(\vec{x}_{I_Q'}) = 0 \quad i \in I' \quad (III.9.4)$$

gelte, so folgt aus (III.9.2) und (III.9.4):

$$\vec{r}_I(\vec{x}_{I_Q} - \vec{x}_{I_Q'}) = 0 \quad i \in I \cap I'. \quad (III.9.5)$$

Aus (III.9.5) und $\text{Rang}(A_I) = \text{Rang}(A_{I'}) = n$ entnimmt man

$$\vec{x}_{I_Q} = \vec{x}_{I_Q'},$$

d.h. unter Berücksichtigung von (II.5.3)

$$r_k(\vec{x}_{I_T}) = 0,$$

was einen Widerspruch mit der Voraussetzung $r_k(\vec{x}_{I_T}) > 0$ darstellt. $\square$

IV. BESTIMMUNG DER MAXIMALEN REFERENZ-
ABWEICHUNG BEI JEDEM SCHRITT. STABILITAET
DER AENDERUNGSFORMEL DER PSEUDOINVERSEN
MATRIX A_I^+ . BEISPIELE.

Durch das Kriterium

$$\frac{y_j}{I_j} = \max_{i \in I} \frac{y_i}{I_i}$$

wird die i_j - Gleichung bestimmt, welche aus der ersten Referenz [I] ausscheidet und die Referenzabweichung der zweiten Referenz [I'] sicher vergrössert.

Weiter wird ein Kriterium entwickelt, welches erlaubt, die maximale Referenzabweichung der Referenz [I'] zu erreichen. Dazu werden die reellen Zahlen $\frac{y_i}{I_i}$, $i \in I$ algebraisch geordnet

und der Einfachheit halber wird angenommen, dass

$$\frac{y_{i_1}}{I_{i_1}} < \dots < \frac{y_{i_\sigma}}{I_{i_\sigma}} < \dots < \frac{y_{i_{n+1}}}{I_{i_{n+1}}}$$

ist.

Nach dem vorangehenden (III.8) Satz liefert h_I^{n+1} des letzten

Quotienten sicherlich eine grössere Referenzabweichung.

Um die maximale Referenzabweichung zu bekommen, untersucht man fortlaufend die Referenzabweichungen

$$h_I^\delta, \delta = i_{n+1}, i_n, \dots, i_\sigma,$$

bis man einen Index i_σ findet, für welchen gilt:

$$h_I^\sigma > h_I^{\sigma+1} > h_I^{\sigma+2} > \dots > h_I^{\sigma+1}$$

und

$$h_I^\sigma > h_I^{\sigma-1}.$$

Das Ziel ist zu zeigen, dass h_I^σ tatsächlich die maximale Referenzabweichung liefert.

Dazu werden wir zunächst Ungleichungen herleiten, die zu den Ungleichungen

$$h_I^\theta > h_I^\sigma \quad i_\theta, i_\sigma \in I$$

äquivalent sind.

IV.1. Satz.

Die Ungleichung

$$h_I^\theta > h_I^\sigma$$

gilt genau dann, wenn

$$\begin{aligned} & \left(\frac{y_{i_\theta}}{I_{i_\theta}} - \frac{y_{i_\sigma}}{I_{i_\sigma}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_I}) \|\vec{I}_I\|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_{n+1}} |f_i| \right) + \right. \\ & \left. + 2 \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_{\sigma+1}}^{i_{n+1}} y_i \operatorname{sgn}(f_i) \right) \right) + 2 \frac{y_{i_\sigma}}{I_{i_\sigma}} \left(\sum_{i=i_{\theta+1}}^{i_\sigma} |f_i| \right) r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\sigma}}{I_{i_\sigma}} \\ & - 2 \left(\sum_{i=i_{\theta+1}}^{i_\sigma} y_i \operatorname{sgn}(f_i) \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\sigma}}{I_{i_\sigma}} \right) > 0, \end{aligned} \quad (IV.1.1)$$

bzw. in äquivalenter Form

$$\begin{aligned} & \left(\frac{y_{i_\theta}}{I_{i_\theta}} - \frac{y_{i_\sigma}}{I_{i_\sigma}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle + r_k(\vec{x}_{I_I}) \|\vec{I}_I\|_1 - 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_1}^{i_\theta} |f_i| \right) - \right. \\ & \left. - 2 \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_1}^{i_\sigma} y_i \operatorname{sgn}(f_i) \right) \right) + 2 \frac{y_{i_\sigma}}{I_{i_\sigma}} \left(\sum_{i=i_{\theta+1}}^{i_\sigma} |f_i| \right) r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\sigma}}{I_{i_\sigma}} \\ & - 2 \left(\sum_{i=i_{\sigma+1}}^{i_\sigma} y_i \operatorname{sgn}(f_i) \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\sigma}}{I_{i_\sigma}} \right) > 0 \end{aligned} \quad (IV.1.2)$$

ist.

Beweis:

Wir zeigen eine Richtung der Behauptung.

Es gelten:

$$\operatorname{sgn}\left(\frac{y_{i_\sigma}}{I_{i_\sigma}} f_i - y_i\right) = \operatorname{sgn}(f_i) \quad i = i_1, \dots, i_\sigma \quad (\text{IV.1.3})$$

$$\operatorname{sgn}\left(\frac{y_{i_\sigma}}{I_{i_\sigma}} f_i - y_i\right) = -\operatorname{sgn}(f_i) \quad i = i_{\sigma+1}, \dots, i_{n+1}. \quad (\text{IV.1.4})$$

Aus $h_I^\theta > h_I^\sigma$ und Formel (III.7.1) folgt:

$$\begin{aligned} & \left(\frac{y_{i_\sigma}}{I_{i_\sigma}} - \frac{y_{i_\sigma}}{I_{i_\sigma}} \right) \langle \vec{r}_I, \vec{b}_I \rangle + \left(r_k(\vec{x}_{I_Q}) + \langle \vec{r}_I, \vec{b}_I \rangle \frac{y_{i_\sigma}}{I_{i_\sigma}} \right) \left(\sum_{i \in I} \left| \frac{y_{i_\sigma}}{I_{i_\sigma}} f_i - y_i \right| \right) - \\ & - \left(r_k(\vec{x}_{I_Q}) + \langle \vec{r}_I, \vec{b}_I \rangle \frac{y_{i_\sigma}}{I_{i_\sigma}} \right) \left(\sum_{i \in I} \left| \frac{y_{i_\sigma}}{I_{i_\sigma}} f_i - y_i \right| \right) > 0. \end{aligned} \quad (\text{IV.1.5})$$

Aus (IV.1.3) und (IV.1.4) folgt:

$$\begin{aligned} \sum_{i \in I} \left| \frac{y_{i_\sigma}}{I_{i_\sigma}} f_i - y_i \right| &= \sum_{i=i_1}^{i_\theta} \frac{y_{i_\sigma}}{I_{i_\sigma}} |f_i| - \sum_{i=i_1}^{i_\theta} y_i \operatorname{sgn}(f_i) - \sum_{i=i_{\theta+1}}^{i_\sigma} \frac{y_{i_\sigma}}{I_{i_\sigma}} |f_i| + \\ &+ \sum_{i=i_{\theta+1}}^{i_\sigma} y_i \operatorname{sgn}(f_i) - \sum_{i=i_{\sigma+1}}^{i_{n+1}} \frac{y_{i_\sigma}}{I_{i_\sigma}} |f_i| + \sum_{i=i_{\sigma+1}}^{i_{n+1}} y_i \operatorname{sgn}(f_i) \end{aligned} \quad (\text{IV.1.6})$$

und

$$\begin{aligned} \sum_{i \in I} \left| \frac{y_{i_\sigma}}{I_{i_\sigma}} f_i - y_i \right| &= \sum_{i=i_1}^{i_\theta} \frac{y_{i_\sigma}}{I_{i_\sigma}} |f_i| - \sum_{i=i_1}^{i_\theta} y_i \operatorname{sgn}(f_i) + \sum_{i=i_{\theta+1}}^{i_\sigma} \frac{y_{i_\sigma}}{I_{i_\sigma}} |f_i| - \\ &- \sum_{i=i_{\theta+1}}^{i_\sigma} y_i \operatorname{sgn}(f_i) - \sum_{i=i_{\sigma+1}}^{i_{n+1}} \frac{y_{i_\sigma}}{I_{i_\sigma}} |f_i| + \sum_{i=i_{\sigma+1}}^{i_{n+1}} y_i \operatorname{sgn}(f_i). \end{aligned} \quad (\text{IV.1.7})$$

Man setzt (IV.1.6), (IV.1.7) in (IV.1.5) ein, so erhält man

$$\begin{aligned}
 & \left(\frac{y_{i_\theta}}{I_{i_\theta}} - \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) \langle \vec{I}_I, \vec{b}_I \rangle + r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_1}^{i_\theta} |f_i| \right) \left(\frac{y_{i_\epsilon}}{I_{i_\epsilon}} - \frac{y_{i_\theta}}{I_{i_\theta}} \right) + \\
 & + r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} |f_i| \right) \left(\frac{y_{i_\theta}}{I_{i_\theta}} + \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) - 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} y_i \operatorname{sgn}(f_i) \right) + \\
 & + r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\epsilon+1}}^{i_{n+1}} |f_i| \right) \left(\frac{y_{i_\theta}}{I_{i_\theta}} - \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) + \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_1}^{i_\theta} y_i \operatorname{sgn}(f_i) \right) \left(\frac{y_{i_\epsilon}}{I_{i_\epsilon}} - \frac{y_{i_\theta}}{I_{i_\theta}} \right) \\
 & + 2 \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\theta}}{I_{i_\theta}} \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} |f_i| \right) - \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_{\epsilon+1}}^{i_\epsilon} y_i \operatorname{sgn}(f_i) \right) \left(\frac{y_{i_\theta}}{I_{i_\theta}} + \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) + \\
 & + \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_{\epsilon+1}}^{i_{n+1}} y_i \operatorname{sgn}(f_i) \right) \left(\frac{y_{i_\theta}}{I_{i_\theta}} - \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) > 0. \quad (\text{IV.1.8})
 \end{aligned}$$

Setzt man

$$\frac{y_{i_\theta}}{I_{i_\theta}} + \frac{y_{i_\epsilon}}{I_{i_\epsilon}} = \frac{y_{i_\theta}}{I_{i_\theta}} - \frac{y_{i_\epsilon}}{I_{i_\epsilon}} + 2 \frac{y_{i_\epsilon}}{I_{i_\epsilon}}$$

in (IV.1.8) ein, dann bekommt man:

$$\begin{aligned}
 & \left(\frac{y_{i_\theta}}{I_{i_\theta}} - \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_1}^{i_\theta} |f_i| \right) + r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_{n+1}} |f_i| \right) - \right. \\
 & \left. - \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_1}^{i_\epsilon} y_i \operatorname{sgn}(f_i) \right) + \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_{\epsilon+1}}^{i_{n+1}} y_i \operatorname{sgn}(f_i) \right) \right) + \\
 & + 2r_k(\vec{x}_{I_Q}) \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} |f_i| \right) - 2 \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} y_i \operatorname{sgn}(f_i) \right) + \\
 & + 2 \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\theta}}{I_{i_\theta}} \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} |f_i| \right) - 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} y_i \operatorname{sgn}(f_i) \right) =
 \end{aligned}$$

$$\begin{aligned}
 &= \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_6}}{I_{i_6}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_Q}) \| \vec{I}_I \|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_{n+1}} |f_i| \right) + \right. \\
 &\quad \left. + \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_1}^{i_{n+1}} y_i \operatorname{sgn}(f_i) \right) - 2 \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_1}^{i_\sigma} y_i \operatorname{sgn}(f_i) \right) \right) + \\
 &\quad + 2 \frac{y_{i_6}}{I_{i_6}} \left(\sum_{i=i_{\theta+1}}^{i_\sigma} |f_i| \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} \right) - \\
 &\quad - 2 \left(\sum_{i=i_{\theta+1}}^{i_\sigma} y_i \operatorname{sgn}(f_i) \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_6}}{I_{i_6}} \right) > 0. \quad (\text{IV.1.9})
 \end{aligned}$$

Da aus (II.5.3)

$$r_k(\vec{x}_{I_I}) = r_k(\vec{x}_{I_Q}) + h_I \sum_{i \in I} y_i \operatorname{sgn}(f_i)$$

ist, folgt die erste Behauptung

$$\begin{aligned}
 &\left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_6}}{I_{i_6}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_I}) \| \vec{I}_I \|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_{n+1}} |f_i| \right) + \right. \\
 &\quad \left. + 2 \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_{\sigma+1}}^{i_{n+1}} y_i \operatorname{sgn}(f_i) \right) \right) + 2 \frac{y_{i_6}}{I_{i_6}} \left(\sum_{i=i_{\theta+1}}^{i_\sigma} |f_i| \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} \right) \\
 &\quad - 2 \left(\sum_{i=i_{\theta+1}}^{i_\sigma} y_i \operatorname{sgn}(f_i) \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_6}}{I_{i_6}} \right) > 0 \quad (\text{IV.1.10})
 \end{aligned}$$

oder durch Einsetzung von

$$2 \langle \vec{I}_I, \vec{b}_I \rangle \sum_{i \in I} y_i \operatorname{sgn}(f_i) = (2r_k(\vec{x}_{I_I}) - 2r_k(\vec{x}_{I_Q})) \| \vec{I}_I \|_1$$

in (IV.1.10) die äquivalente Ungleichung

$$\begin{aligned} & \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_e}}{I_{i_e}} \right) \left(\langle \vec{T}_I, \vec{b}_I \rangle + r_k(\vec{x}_{I_T}) \|\vec{T}_I\|_1 - 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=1, i \neq i_e}^{i_0} |f_i| \right) - \right. \\ & \left. - 2 \langle \vec{T}_I, \vec{b}_I \rangle \left(\sum_{i=i_1}^{i_e} y_i \operatorname{sgn}(f_i) \right) \right) + 2 \frac{y_{i_e}}{I_{i_e}} \left(\sum_{i=i_{e-1}}^{i_e} |f_i| \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{T}_I, \vec{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} \right) - \\ & - 2 \left(\sum_{i=i_{\theta+1}}^{i_e} y_i \operatorname{sgn}(f_i) \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{T}_I, \vec{b}_I \rangle \frac{y_{i_e}}{I_{i_e}} \right) > 0. \end{aligned} \quad (\text{IV.1.11})$$

Die Umkehrung ist offensichtlich. $\square$

IV.2. Satz.

Ist

$$h_I^{\sigma-1} > h_I^{\sigma},$$

so gilt

$$\begin{aligned} & \langle \vec{T}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_T}) \|\vec{T}_I\|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_e}^{i_{\theta+1}} |f_i| \right) + \\ & + 2 \langle \vec{T}_I, \vec{b}_I \rangle \left(\sum_{i=i_e}^{i_{\theta+1}} y_i \operatorname{sgn}(f_i) \right) < 0. \end{aligned} \quad (\text{IV.2.1})$$

Beweis:

Im 2. Teil von (IV.1.10) ist für $\theta = \sigma - 1$

$$\begin{aligned} & 2 \frac{y_{i_e}}{I_{i_e}} |f_{i_e}| \left(r_k(\vec{x}_{I_Q}) + \langle \vec{T}_I, \vec{b}_I \rangle \frac{y_{i_{\sigma+1}}}{I_{i_{\sigma+1}}} - 2 y_{i_e} \operatorname{sgn}(f_{i_e}) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{T}_I, \vec{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} \right) \right) \\ & = 2 \langle \vec{T}_I, \vec{b}_I \rangle y_{i_e} \operatorname{sgn}(f_{i_e}) \left(\frac{y_{i_{\sigma+1}}}{I_{i_{\sigma+1}}} - \frac{y_{i_0}}{I_{i_0}} \right) \end{aligned} \quad (\text{IV.2.2})$$

Setzt man (IV.2.2) in (IV.1.10) ein, so ergibt sich die Ungleichung

$$\begin{aligned} & \left(\frac{y_{i_{\sigma+1}}}{I_{i_{\sigma+1}}} - \frac{y_{i_e}}{I_{i_e}} \right) \left(\langle \vec{T}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_T}) \|\vec{T}_I\|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_e}^{i_{\theta+1}} |f_i| \right) + \right. \\ & \left. + 2 \langle \vec{T}_I, \vec{b}_I \rangle \left(\sum_{i=i_e}^{i_{\theta+1}} y_i \operatorname{sgn}(f_i) \right) \right) > 0. \end{aligned}$$

Da $\frac{y_{i_{\sigma}}}{I_{i_{\sigma}}} - \frac{y_{i_{\sigma}}}{I_{i_{\sigma}}} < 0$ ist, so folgt die Behauptung (IV.2.1). Die

Umkehrung ist offensichtlich. $\square$

IV.3. Korollar.

Aus

$$h_I^n > h_I^{n+1}$$

ergibt sich nach (IV.2.1)

$$\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_T}) \|\vec{I}_I\|_1 + |I_{i_{n+1}}| (r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}}) < 0, \quad (IV.3.1)$$

Mit Hilfe der Sätze IV.1, IV.2 und des Korollars IV.3 lässt sich beweisen, dass falls $h_I^n > h_I^{n+1}$ ist, so ist h_I^{n+1} die maximale Referenzabweichung.

IV.4. Satz.

Ist

$$h_I^{n+1} > h_I^n,$$

so ist

$$h_I^{n+1} = \max_{i \in I} h_I^i.$$

Beweis:

Man formt die linke Seite der Ungleichung (IV.1.9) für $i_{\sigma} = i_{n+1}$,

$i_{\theta} \in I$, $i_{\theta} \neq i_{n+1}$ folgendermassen um:

$$\begin{aligned} & \left(\frac{y_{i_{\theta}}}{I_{i_{\theta}}} - \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_Q}) \|\vec{I}_I\|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_{n+1}} |I_i| \right) - \right. \\ & \left. - \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_{\theta}}^{i_{n+1}} y_i \operatorname{sgn}(I_i) \right) \right) + 2 \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \left(\sum_{i=i_{\theta+1}}^{i_{n+1}} |I_i| \right) (r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}}) + \\ & + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{\theta}}}{I_{i_{\theta}}} - 2 \left(\sum_{i=i_{\theta+1}}^{i_{n+1}} y_i \operatorname{sgn}(I_i) \right) (r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}}) = \end{aligned}$$

$$\begin{aligned}
 &= \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_I}) \|\vec{I}_I\|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{0+1}}^{i_{n+1}} |f_i| \right) \right) + \\
 &\quad + 2 \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \left(\sum_{i=i_{0+1}}^{i_{n+1}} |f_i| \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} - 2 \left(\sum_{i=i_{0+1}}^{i_{n+1}} y_i \operatorname{sgn}(f_i) \right) \right) r_k(\vec{x}_{I_Q}) \\
 &\quad + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \Bigg| = \\
 &= \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_I}) \|\vec{I}_I\|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{0+1}}^{i_{n+1}} |f_i| \right) \right) + \\
 &\quad + 2 \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \right) \left(\sum_{i=i_{0+1}}^{i_n} |f_i| \right) r_k(\vec{x}_{I_Q}) + 2 \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \left(\sum_{i=i_{0+1}}^{i_n} |f_i| \right) r_k(\vec{x}_{I_Q}) + \\
 &\quad + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} \Bigg) + 2 \frac{y_{i_{n+1}}}{I_{i_{n+1}}} |f_{i_{n+1}}| r_k(\vec{x}_{I_Q}) + 2 \frac{y_{i_{n+1}}}{I_{i_{n+1}}} |f_{i_{n+1}}| \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} - \\
 &\quad - 2 \left(\sum_{i=i_{0+1}}^{i_n} y_i \operatorname{sgn}(f_i) \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}} - 2 y_{i_{n+1}} \operatorname{sgn}(f_{i_{n+1}}) r_k(\vec{x}_{I_Q}) - \right. \\
 &\quad \left. - 2 y_{i_{n+1}} \operatorname{sgn}(f_{i_{n+1}}) \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \right) = \\
 &= \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_I}) \|\vec{I}_I\|_1 + 2 |f_{i_{n+1}}| r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \right) + \\
 &\quad + 2 \frac{y_{i_0}}{I_{i_0}} \left(\sum_{i=i_{0+1}}^{i_n} |f_i| \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \right) - 2 \left(\sum_{i=i_{0+1}}^{i_n} y_i \operatorname{sgn}(f_i) \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \right) + \\
 &\quad + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \Bigg| =
 \end{aligned}$$

$$= \left(\frac{y_{i_\theta}}{f_{i_\theta}} - \frac{y_{i_{n+1}}}{f_{i_{n+1}}} \right) \left(\langle \vec{f}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_T}) \|\vec{f}_I\|_1 + 2 |f_{i_{n+1}}| (r_k(\vec{x}_{I_Q}) + \langle \vec{f}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{f_{i_{n+1}}}) \right) + 2 (r_k(\vec{x}_{I_Q}) + \langle \vec{f}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{f_{i_{n+1}}}) \left(\sum_{i=i_{\theta+1}}^{i_n} \left(\frac{y_{i_\theta}}{f_{i_\theta}} - \frac{y_i}{f_i} \right) |f_i| \right).$$

Der letzte Ausdruck ist < 0 ,

weil

$$a) \frac{y_{i_\theta}}{f_{i_\theta}} - \frac{y_{i_{n+1}}}{f_{i_{n+1}}} < 0 \quad \text{für } \forall i_\theta \in I, \quad i_\theta \neq i_{n+1}$$

$$b) \langle \vec{f}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_T}) \|\vec{f}_I\|_1 + 2 |f_{i_{n+1}}| (r_k(\vec{x}_{I_Q}) + \langle \vec{f}_I, \vec{b}_I \rangle \frac{y_i}{f_i}) > 0$$

(nach Korollar IV.3),

$$c) r_k(\vec{x}_{I_Q}) + \langle \vec{f}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{f_{i_{n+1}}} = r_k(\vec{x}_{I_T}) - h_I \sum_{i \in I} y_i \operatorname{sgn}(f_i) +$$

$$+ \frac{y_{i_{n+1}}}{f_{i_{n+1}}} h_I \|\vec{f}_I\|_1 = r_k(\vec{x}_{I_T}) + h_I \left(\sum_{i=i_\theta}^{i_n} \left(\frac{y_{i_{n+1}}}{f_{i_{n+1}}} - \frac{y_i}{f_i} \right) |f_i| \right) > 0,$$

$$\text{da } r_k(\vec{x}_{I_T}) > 0, \quad h_I > 0 \text{ und } \frac{y_{i_{n+1}}}{f_{i_{n+1}}} - \frac{y_i}{f_i} > 0$$

$$i=i_1, \dots, i_{n+1},$$

$$d) \frac{y_{i_\theta}}{f_{i_\theta}} - \frac{y_i}{f_i} < 0 \quad \text{für } i=i_{\theta+1}, i_{\theta+2}, \dots, i_n \quad \text{sind.} \quad \square$$

IV.5. Satz.

Wenn für einen Index $i_\theta \in I$

$$h_I^\sigma > h_I^{\sigma-1}$$

und

$$h_I^\sigma > h_I^{\sigma+1} > h_I^{\sigma+2} > \dots > h_I^{n+1}$$

gilt, dann ist

$$h_I^\sigma = \max_{i \in I} h_I^i.$$

Beweis:

Man formt die linke Seite der Ungleichung (IV.1.1) für $i_1 \leq i_\theta \leq i_{\sigma-1}$ folgendermassen um:

$$\begin{aligned}
 & \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_I}) \|\vec{I}_I\|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=1}^{i_{n+1}} |f_i| \right) + \right. \\
 & \quad \left. + 2\langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=1}^{i_{n+1}} y_i \operatorname{sgn}(f_i) \right) \right) + \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) \left(2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_{n+1}} |f_i| \right) - \right. \\
 & \quad \left. - 2\langle \vec{I}_I, \vec{b}_I \rangle y_{i_\epsilon} \operatorname{sgn}(f_{i_\epsilon}) \right) + 2 \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} |f_i| \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} \right) - \\
 & \quad - 2 \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} y_i \operatorname{sgn}(f_i) \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) = \\
 & = \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_I}) \|\vec{I}_I\|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_\epsilon}^{i_{n+1}} |f_i| \right) + \right. \\
 & \quad \left. + 2\langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_\epsilon}^{i_{n+1}} y_i \operatorname{sgn}(f_i) \right) \right) + \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) \left(-2\langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} |f_{i_\epsilon}| + \right. \\
 & \quad \left. + 2 \frac{y_{i_0}}{I_{i_0}} r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_{n+1}} |f_i| \right) - 2 \frac{y_{i_\epsilon}}{I_{i_\epsilon}} r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_{n+1}} |f_i| \right) + \right. \\
 & \quad \left. + 2r_k(\vec{x}_{I_Q}) \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \left(\sum_{i=i_{\theta+1}}^{i_{\epsilon-1}} |f_i| \right) + 2 \frac{y_{i_\epsilon}}{I_{i_\epsilon}} |f_{i_\epsilon}| r_k(\vec{x}_{I_Q}) + 2 \frac{y_{i_\epsilon} y_{i_0}}{I_{i_\epsilon} I_{i_0}} \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} |f_i| \right) \right. \\
 & \quad \left. - 2 \left(\sum_{i=i_{\theta+1}}^{i_\epsilon} y_i \operatorname{sgn}(f_i) \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) - 2 \frac{y_{i_\epsilon}}{I_{i_\epsilon}} |f_{i_\epsilon}| \left(r_k(\vec{x}_{I_Q}) + \right. \right. \\
 & \quad \left. \left. + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) \right) = \\
 & \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) \left(\langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_{I_I}) \|\vec{I}_I\|_1 + 2r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_\epsilon}^{i_{n+1}} |f_i| \right) + \right. \\
 & \quad \left. + 2\langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_\epsilon}^{i_{n+1}} y_i \operatorname{sgn}(f_i) \right) \right) + 2 \frac{y_{i_\epsilon} y_{i_0}}{I_{i_\epsilon} I_{i_0}} \langle \vec{I}_I, \vec{b}_I \rangle \left(\sum_{i=i_{\theta+1}}^{i_{\epsilon-1}} |f_i| \right) + \\
 & \quad + 2 \frac{y_{i_0}}{I_{i_0}} r_k(\vec{x}_{I_Q}) \left(\sum_{i=i_{\theta+1}}^{i_{\epsilon-1}} |f_i| \right) - 2 \left(\sum_{i=i_{\theta+1}}^{i_{\epsilon-1}} y_i \operatorname{sgn}(f_i) \right) \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\epsilon}}{I_{i_\epsilon}} \right) =
 \end{aligned}$$

$$\begin{aligned}
 &= \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_0}}{I_{i_0}} \right) \left(\langle \mathbf{I}_I, \mathbf{b}_I \rangle - r_k(\tilde{x}_{I_I}) \|\mathbf{I}_I\|_1 + 2r_k(\tilde{x}_{I_Q}) \left(\sum_{i=i_0}^{i_{n+1}} |I_i| \right) + \right. \\
 &\quad \left. + 2\langle \mathbf{I}_I, \mathbf{b}_I \rangle \left(\sum_{i=i_0}^{i_{n+1}} y_i \operatorname{sgn}(I_i) \right) \right) + 2 \frac{y_{i_0}}{I_{i_0}} \left(\sum_{i=i_0}^{i_{n+1}} |I_i| \right) r_k(\tilde{x}_{I_Q}) + \langle \mathbf{I}_I, \mathbf{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} \Bigg) - \\
 &\quad - 2 \left(\sum_{i=i_{n+1}}^{i_0} y_i \operatorname{sgn}(I_i) \right) \left(r_k(\tilde{x}_{I_Q}) + \langle \mathbf{I}_I, \mathbf{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} \right) = \\
 &= \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_0}}{I_{i_0}} \right) \left(\langle \mathbf{I}_I, \mathbf{b}_I \rangle - r_k(\tilde{x}_{I_I}) \|\mathbf{I}_I\|_1 + 2r_k(\tilde{x}_{I_Q}) \left(\sum_{i=i_0}^{i_{n+1}} |I_i| \right) + \right. \\
 &\quad \left. + 2\langle \mathbf{I}_I, \mathbf{b}_I \rangle \left(\sum_{i=i_0}^{i_{n+1}} y_i \operatorname{sgn}(I_i) \right) \right) + 2 \left(r_k(\tilde{x}_{I_Q}) + \langle \mathbf{I}_I, \mathbf{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} \right) \left(\sum_{i=i_{n+1}}^{i_0} \left(\frac{y_{i_0}}{I_{i_0}} - \frac{y_i}{I_i} \right) |I_i| \right).
 \end{aligned}$$

Der letzte Ausdruck ist < 0 ,

weil

$$a) \frac{y_{i_0}}{I_{i_0}} - \frac{y_{i_0}}{I_{i_0}} < 0 \quad \text{für} \quad i_0 \in I, \quad i_0 = i_1, \dots, i_{n-1},$$

$$\begin{aligned}
 b) \quad &\langle \mathbf{I}_I, \mathbf{b}_I \rangle - r_k(\tilde{x}_{I_I}) \|\mathbf{I}_I\|_1 + 2r_k(\tilde{x}_{I_Q}) \sum_{i=i_0}^{i_{n+1}} |I_i| + \\
 &+ 2\langle \mathbf{I}_I, \mathbf{b}_I \rangle \left(\sum_{i=i_0}^{i_{n+1}} y_i \operatorname{sgn}(I_i) \right) > 0
 \end{aligned}$$

(nach Satz (IV.2.)) ,

$$c) r_k(\tilde{x}_{I_Q}) + \langle \mathbf{I}_I, \mathbf{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} > 0.$$

Denn nach Satz (IV.2.) ist die Voraussetzung

$$h_I^\sigma > h_I^{\sigma+1} > \dots, > h_I^{n+1}$$

äquivalent zu der Bedingung

$$\begin{aligned}
 &\langle \mathbf{I}_I, \mathbf{b}_I \rangle - r_k(\tilde{x}_{I_I}) \|\mathbf{I}_I\|_1 + |I_{i_{n+1}}| \left(r_k(\tilde{x}_{I_Q}) + \langle \mathbf{I}_I, \mathbf{b}_I \rangle \frac{y_{i_{n+1}}}{I_{i_{n+1}}} \right) + \\
 &+ \dots + |I_{i_0}| \left(r_k(\tilde{x}_{I_Q}) + \langle \mathbf{I}_I, \mathbf{b}_I \rangle \frac{y_{i_0}}{I_{i_0}} \right) < 0
 \end{aligned}$$

und die Voraussetzung

$$h_I^\sigma > h_I^{\sigma-1}$$

äquivalent zu der Bedingung

$$\begin{aligned} & \langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_I) \parallel \vec{I}_I \parallel_1 + |\vec{I}_{i_{n+1}}| \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{\vec{I}_{i_{n+1}}} \right) + \dots + \\ & + |\vec{I}_{i_{\epsilon+1}}| \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{\epsilon+1}}}{\vec{I}_{i_{\epsilon+1}}} \right) + |\vec{I}_{i_\epsilon}| \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\epsilon}}{\vec{I}_{i_\epsilon}} \right) > 0 \end{aligned}$$

Man hat nur die eine Richtung der Sätze (IV.4.) (IV.5)

bewiesen, weil die Gültigkeit der anderen Richtung offensichtlich ist. $\square$

IV.6. Kriterium

Man ordnet algebraisch die reellen Zahlen $\frac{y_{i_i}}{\vec{I}_{i_i}} \quad i_i \in I$ und

untersucht sie fortlaufend in abnehmender Richtung bis man einen Index $j := i_\sigma$ findet für welchen die Bedingungen

$$\begin{aligned} & \langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_I) \parallel \vec{I}_I \parallel_1 + |\vec{I}_{i_{n+1}}| \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{\vec{I}_{i_{n+1}}} \right) + \dots + \\ & + 2 |\vec{I}_{i_{\epsilon+1}}| \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{\epsilon+1}}}{\vec{I}_{i_{\epsilon+1}}} \right) < 0 \end{aligned}$$

und

$$\begin{aligned} & \langle \vec{I}_I, \vec{b}_I \rangle - r_k(\vec{x}_I) \parallel \vec{I}_I \parallel_1 + |\vec{I}_{i_{n+1}}| \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{n+1}}}{\vec{I}_{i_{n+1}}} \right) + \dots + \\ & + 2 |\vec{I}_{i_{\epsilon+1}}| \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_{\epsilon+1}}}{\vec{I}_{i_{\epsilon+1}}} \right) + |\vec{I}_{i_\epsilon}| \left(r_k(\vec{x}_{I_Q}) + \langle \vec{I}_I, \vec{b}_I \rangle \frac{y_{i_\epsilon}}{\vec{I}_{i_\epsilon}} \right) > 0 \end{aligned}$$

gelten. Nach (IV.2), (IV.4.) und (IV.5.) ist $h_I^j := h_I^\sigma$ die maximale Referenzabweichung, d.h.

$$h_I^j = \max_{i \in I} h_I^i$$

IV.7. Fehlerabschätzung der pseudoinversen Matrix A_I^+ .

Man betrachtet nun den Fehler, der bei der Berechnung von A_I^+ auftritt.

Unter Benützung der Gleitpunktarithmetik ergibt sich das berechnete

$\vec{a}^{i+}$ wie folgt:

$$\vec{a}^{i+} = \text{gl}(\vec{a}^{i+} - y_i \vec{z} - f_i \vec{z}') \quad i \in I / \{j\} \quad j \in I$$

oder

$$\hat{a}_{ti}^+ = \text{gl}(a_{ti}^+ - y_i z_t - f_i z'_t), \quad (\text{IV.7.1})$$

wobei

$$z_t = \frac{1}{uv} (f_j^2 d_t + y_j a_{tj}^+) \quad (\text{IV.7.1})$$

$$z_t = \frac{f_j}{lv} (v a_{tj}^+ - y_j d_t) \quad (t=1, \dots, n) \quad (\text{IV.7.3})$$

und a_{ti}^+ , bzw. a_{ti}^+ das (t,i) Element der Matrix A_I^+ , bzw. A_I^+ , d_t der t -Komponent des Vektors $\vec{d}$ ist.

Unter Berücksichtigung von (I.13.1-I.13.2) gelten:

$$\text{gl}(y_i f_j^2 d_t) = y_i f_j^2 d_t (1 + \epsilon_1) \quad |\epsilon_1| \leq 2^{-t+1} \quad (\text{IV.7.4})$$

$$\text{gl}(y_i y_j a_{tj}^+) = y_i y_j a_{tj}^+ (1 + \epsilon_2) \quad |\epsilon_2| \leq 2^{-t} \sqrt{3} \quad (\text{IV.7.5})$$

$$\text{gl}(f_i f_j y_j d_t) = f_i f_j y_j d_t (1 + \epsilon_3) \quad |\epsilon_3| \leq 2^{-t+1} \quad (\text{IV.7.6})$$

$$\text{gl}(f_i f_j v a_{tj}^+) = f_i f_j v a_{tj}^+ (1 + \epsilon_4) \quad |\epsilon_4| \leq 2^{-t+1}. \quad (\text{IV.7.7})$$

Aus (I.12.1-I.12.4) und (IV.7.1) folgt:

$$\begin{aligned}
\tilde{a}_{ti}^+ &= \left(a_{ti}^+ - \frac{1}{\mu\nu} \left[(y_i f_j^2 d_t (1+\epsilon_1) + y_i y_j a_{tj}^+ (1+\epsilon_2)) (1+\rho_1) \right] (1+\delta) \right) (1+\rho_2) - \\
&\quad \frac{1}{\mu\nu} \left[(f_i f_j v a_{tj}^+ (1+\epsilon_3) - f_i f_j y_j d_t (1+\epsilon_3)) (1+\rho_2) \right] (1+\delta) (1+\rho_3) = \\
&= a_{ti}^+ (1+\rho_2)(1+\rho_3) - \frac{1}{\mu\nu} y_i f_j^2 d_t (1+\epsilon_1)(1+\rho_1)(1+\delta)(1+\rho_2)(1+\rho_3) - \\
&\quad - \frac{1}{\mu\nu} y_i y_j a_{tj}^+ (1+\epsilon_2)(1+\rho_1)(1+\delta)(1+\rho_2)(1+\rho_3) - \\
&\quad - \frac{1}{\mu\nu} f_i f_j v a_{tj}^+ (1+\epsilon_3)(1+\rho_2)(1+\delta)(1+\rho_3) + \\
&\quad + \frac{1}{\mu\nu} f_i f_j y_j d_t (1+\epsilon_3)(1+\rho_2)(1+\delta)(1+\rho_3) = \\
&= a_{ti}^+ (1+\gamma) - \frac{1}{\mu\nu} y_i f_j^2 d_t (1+\gamma_2) - \frac{1}{\mu\nu} y_i y_j a_{tj}^+ (1+\gamma_3) - \frac{1}{\mu\nu} f_i f_j v a_{tj}^+ (1+\gamma_4) \\
&\quad + \frac{1}{\mu\nu} f_i f_j y_j d_t (1+\gamma_4) \quad (\text{IV.7.8})
\end{aligned}$$

mit

$$(1-2^{-t})^2 \leq 1+\gamma_1 \leq (1+2^{-t})^2 \quad (\text{IV.7.9})$$

$$(1-2^{-t})^4 \leq 1+\gamma_2 \leq (1+2^{-t})^4 \quad (\text{IV.7.10})$$

$$(1-2^{-t})^6 \leq 1+\gamma_3 \leq (1+2^{-t})^6 \quad (\text{IV.7.11})$$

Aus (I4.5) und (IV.7.8-IV.7.11) folgt:

$$\begin{aligned}
|\epsilon_{ti}| &= |\tilde{a}_{ti}^+ - a_{ti}^+| \leq |\tilde{a}_{ti}^+ \gamma_1| + \frac{1}{\mu\nu} |y_i f_j^2 d_t| + |y_i y_j a_{tj}^+| (5 \times 2^{-t_1}) + \\
&\quad + \frac{1}{\mu\nu} |f_i f_j v a_{tj}^+| + |f_i f_j y_j d_t| (4 \times 2^{-t_1})
\end{aligned}$$

oder

$$\begin{aligned}
 |c_{ti}| &\leq 2^{-t_i+1} \left(|a_{ti}^+| + \frac{2.5}{\mu\nu} (f_j^2 |y_i d_t| + |y_i y_j| |a_{tj}^+|) \right) + \\
 &+ 2^{-t_i+1} \left(\frac{2}{\mu\nu} (|y_j f_i f_j d_t| + \nu |f_j f_i a_{tj}^+|) \right) \leq \\
 &\leq 2^{-t_i+1} \left(|a_{ti}^+| + \frac{2.5}{\mu\nu} (f_j^2 |y_i y_\zeta| \|A_I^+\|_\infty + |y_i y_j a_{tj}^+|) \right) + \\
 &+ \frac{2}{\mu\nu} (|y_j f_i f_j y_\zeta| \|A_I^+\|_\infty + \nu |f_j f_i a_{tj}^+|) \quad , \quad (IV.7.12)
 \end{aligned}$$

da

$$|d_t| \leq \sum_{i=1}^{n+1} |a_{ti}^+ y_i| \leq |y_\zeta| \left(\sum_{i=1}^{n+1} |a_{ti}^+| \right) \leq |y_\zeta| \|A_I^+\|_\infty$$

mit

$$|y_\zeta| = \max_{1 \leq i \leq n+1} |y_i|$$

ist.

Es gelten:

$$a) \frac{1}{\mu\nu} |f_j^2 y_i y_\zeta| = \frac{f_j^2 |y_i y_\zeta|}{f_j^2 (1 + \|\vec{y}\|_2^2) + y_j} < 1 \quad , \quad (IV.7.13)$$

$$b) \frac{1}{\mu\nu} |y_\zeta y_j f_i f_j| = \frac{|1_i|}{\frac{|f_j|}{|y_j y_\zeta|} + \frac{|f_j y_j|}{|y_\zeta|} + \frac{|f_j y_\zeta|}{|y_j|} + \frac{|y_j|}{|f_j y_\zeta|} + \frac{|f_j|}{|y_j y_\zeta|} \left(\sum_{i \neq j, \zeta}^{n+1} y_i^2 \right)} < 1 \quad ,$$

(IV.7.14)

c) Aus den offensichtlichen Ungleichungen

$$\frac{|y_j|}{\sqrt{\mu\nu}} < 1, \quad \frac{|f_j|}{\sqrt{\mu}} < 1 \quad \text{und} \quad \frac{|y_i|}{\sqrt{\nu}} < 1$$

ergibt sich

$$\frac{y_1 y_j + f_1 f_j \nu}{\mu\nu} \cdot \frac{|y_i|}{\sqrt{\mu\nu}} + \frac{|f_i|}{\sqrt{\mu}} = \frac{\epsilon_0 + |f_i|}{\sqrt{\mu}}, \quad \epsilon_0 := \frac{|y_i|}{\sqrt{\nu}} < 1 \quad (\text{IV.7.15}).$$

Aus (IV.7.12-IV.7.15) folgt:

$$|\epsilon_{ti}| = |\tilde{a}_{ti}^+ - a_{ti}^+| < 2^{-t_1+1} \left(5 + 2.5 \frac{\epsilon_0 + |f_i|}{\sqrt{\mu}} \right) \|A_I^+\|_\infty.$$

Falls $|f_i| \leq \sqrt{\mu}$ d.h. $f_i^2 \leq f_j^2 + \frac{y_j^2}{\nu}$ ist,

gilt:

$$\frac{\epsilon_0 + |f_i|}{\sqrt{\mu}} < 2,$$

andernfalls deuten alle behandelten Beispielen darauf hin, dass

$$\frac{\epsilon_0 + |f_i|}{\sqrt{\mu}} < 2n$$

wird, wobei $n = \text{Rang}(A_I)$ ist. $\square$

IV.8. Iterative Verbesserung der berechneten Matrix $\tilde{A}_{I^*}^+$.

Es sei $A_{I^*}^+$ die exakte Pseudoinverse von A_{I^*} und $\tilde{A}_{I^*}^+$ die berechnete Matrix, welche von der exakten Matrix $A_{I^*}^+$ um eine Fehlermatrix F abweicht:

$$\tilde{A}_{I^*}^+ = A_{I^*}^+ + F. \quad (\text{IV.8.1})$$

Es wird angenommen, dass die Voraussetzungen

$$\rho(FA_{I^*}) < 1 \quad (\text{IV.8.2})$$

und

$$\tilde{A}_{I^*}^+ A_{I^*} A_{I^*}^+ = \tilde{A}_{I^*}^+$$

gelten.

Aus (IV.8.1) folgt:

$$\tilde{A}_{I'}^+ A_{I'} = I_n + F A_{I'} = I_n + D.$$

Weiter sei

$$F = D \tilde{A}_{I'}^+$$

eine Annäherung für F . Verbessert man mit den abgeschätzten F die Matrix $\tilde{A}_{I'}^+$, so ergibt sich

$$\tilde{A}_2^+ = \tilde{A}_{I'}^+ - D \tilde{A}_{I'}^+ = (I_n - D) \tilde{A}_{I'}^+, \quad (\text{IV.8.4})$$

$$\begin{aligned} \tilde{A}_2^+ A_{I'} &= \tilde{A}_{I'}^+ A_{I'} - D \tilde{A}_{I'}^+ A_{I'} = (I_n + D) - D(I_n + D) = \\ &= I_n - D^2, \end{aligned}$$

analog

$$\tilde{A}_3^+ = \tilde{A}_2^+ + D^2 \tilde{A}_2^+ = (I_n + D^2) \tilde{A}_2^+$$

$$\tilde{A}_3^+ A_{I'} = (I_n + D^2) \tilde{A}_2^+ A_{I'} = (I_n + D^2)(I_n - D^2) = I_n - D^4.$$

Allgemein erhält man

$$\tilde{A}_{i+1}^+ = \begin{pmatrix} I_n + (-1)^{2^{i-1}} & D^{2^{i-1}} \end{pmatrix} \tilde{A}_i^+, \quad \tilde{A}_1^+ = \tilde{A}_{I'}^+ \quad i=1,2,\dots$$

und somit

$$\begin{aligned} \tilde{A}_{i+1}^+ &= (I_n - D)(I_n + D^2) \dots (I_n + D^{2^{i-1}}) \tilde{A}_{I'}^+ = \\ &= (I_n - D + D^2 - D^3 + \dots + (-1)^{2^i - 1} D^{2^i - 1}) \tilde{A}_{I'}^+. \end{aligned}$$

Aus (IV.8.2) und (IV.8.3) folgt:

$$\lim_{l \rightarrow \infty} A_{l+1}^+ = (I_n + D)^{-1} (I_n + D) A_l^+ = A_l^+$$

Das Problem der Nachiteration behandelt Sauter [26], wobei er für das Konvergieren der Iteration $X_n = X_{n-1} + X_0(I - AX_{n-1})$

($X_0 = A^+ - F$ ist eine beliebige Näherungslösung von A^+) gegen A^+ die gleichen Voraussetzungen (IV.8.2), (IV.8.3), d.h.

$X_0 A A^+ = X_0$ und $\rho(AF) < 1$ angibt.

Man sieht, dass das vorgeschlagene Iterationsverfahren eine Übertragung der üblichen Iterationsverfahren zur Matrixinversion darstellt, und die Voraussetzung (IV.8.2) entspricht der Voraussetzung für die Konvergenz des Iterationsverfahrens gegen die gesuchte Inverse.

IV. 9. Beschreibung des Algorithmus.

- Start:
- a) man wählt irgendeine Referenz [I] des Systems (II.1.1),
 - b) man berechnet die pseudoinverse Matrix A_I^+ nach (I.9.d), sowie
 - c) $\vec{x}_{I_Q} = A_I^+ \vec{b}_I$,
 - d) $\vec{f}_I$ (nach III.1.17-III.1.18).

Iterationsverfahren:

1. Schritt: Man berechnet:

- a) $h_I = \frac{\langle \vec{f}_I, \vec{b}_I \rangle}{\|\vec{f}_I\|_1}$,
- b) $\vec{x}_{I_T} = \vec{x}_{I_Q} - h_I A_I^+ \text{sgn}(\vec{f}_I)$.

2. Schritt: Man sucht dasjenige k aus, für welches gilt:

$$r_k(\vec{x}_{I_T}) = \max_{i \in I_0} |r_i(\vec{x}_{I_T})|$$

Ist $|r_k(\vec{x}_{I_T})| > |h_I|$, geht man zu Schritt 3 über.

Ist $|r_k(\vec{x}_{I_T})| \leq |h_I|$, so ist die Iteration beendet.

3. Schritt: Man berechnet:

$$a) \vec{y}^T = \text{sgn}(r_k(\vec{x}_{I_I})) \vec{a}_{kI}^T A_{I_I}^T.$$

$$b) \text{ man bildet } \frac{y_i}{I_i} \quad \text{für } i \in I,$$

c) man ordnet algebraisch die $\frac{y_i}{I_i} \quad i \in I$ und sucht dasjenige $j := i_0 \in I$, für welches die Bedingungen des Kriteriums (IV.6) gelten.

4. Schritt: Man bestimmt:

a) die pseudoinverse Matrix $A_{I_I}^+$ nach (III.2.1-III.2.8),

b) den Vektor $\vec{I}_{I_I}$ nach (III.5.1-III.5.2),

c) den Vektor $\vec{x}_{I_I Q}$ nach (III.6.1).

5. Schritt:

a) man setzt $I' = \{I \cup \{k\} / \{j\} \}$

b) $I = I'$

IV.10. Beispiele:

Mit der vorgeschlagenen Methode sowie mit der 2. Methode von Cheney [8] und der von Barrodale I. — Phillips C. [1] wurden die folgenden Beispiele gelöst:

$$a)^* \vec{r}(\vec{x}) = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 2 \\ 2 \\ 4 \end{bmatrix} - \begin{bmatrix} 0 & -1 & 0 \\ 1 & 3 & -4 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$b)^* \vec{r}(\vec{x}) = \begin{bmatrix} 0 \\ -2 \\ 1 \\ -1 \\ 5 \\ 7 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$c_{1,2})^* \vec{r}(\vec{x}) = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 1-\epsilon \\ 1 & 1 \\ 1 & 1-2\epsilon \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$c_1 \text{ für } = .2 \times 10^{-3}$
 $c_2 \text{ für } = .1 \times 10^{-4}$

* a, b stammen aus [10], $c_{1,2}$ aus [22].

d_{1,2,3}) Die Funktion $f(x) = \frac{1}{1+25x^2}$ ist im T-Sinne

durch die Polynome

$$P_{2n}(x) = \sum_{i=0}^n \alpha_i x^{2i} \quad \begin{array}{ll} d_1 & \text{für } n=5 \\ d_2 & \text{für } n=7 \\ d_3 & \text{für } n=10 \end{array}$$

auf dem äquidistanten Gitter

$$G = \{x_i\}_{i=0}^{50}, \quad x_i = \frac{2i}{100} \quad i=0,1,2,\dots,50$$

zu approximieren.

e_{1,2}) Die Funktion $f(x)$ des Beispiels d ist im T-Sinne mit dem Polynom

$$P_{20}(x) = \sum_{i=0}^{10} \alpha_i x^{2i},$$

bzw. mit dem Polynom

$$P_{20}(x) = \sum_{i=0}^{10} \beta_i T_{2i}(x)$$

zu approximieren, wo $T_n(x)$ das n-te T-Polynom darstellt. Als Stützstellen werden für beide Fälle die Punkte

$$x_i = \frac{i}{100} \quad i = 0,1,2,\dots,100$$

genommen.

Das T-Polynom $T_n(x)$ ist durch die Rekursionsformel

$$T_0(x) = 1$$

$$T_1(x) = xT_0(x)$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n \geq 1$$

gegeben.

Die Ausführung der Lösung aller oben erwähnten Beispiele wurde im RZU (Rechenzentrum der Universität Zürich) durchgeführt. Für den Algorithmus von Barrodale und Phillips [1] wurde das FORTRAN IV Programm: ALGORITHM 495, SOLUTION OF AN OVERDETERMINED SYSTEM OF LINEAR EQUATIONS IN THE CHEBYSHEV NORM F4, ACM Transactions on Mathematical Software, Vol. 1, No. 3, September 1975, pages 264-270, von I. Barrodale und C. Phillips benutzt. Der vorgeschlagene Algorithmus und der von Cheney [8] sind in FORTRAN IV programmiert, dabei wurden von der Programmbibliothek der RZU folgende Programme verwendet:

1. Prozedur LIN2F zur Invertierung einer Matrix.
2. Prozedur LPSDOR zur Berechnung der Pseudoinversen einer Matrix.
3. Prozedur LSVDF zur Berechnung der singulären Werte einer Matrix.
4. Prozedur GETTM,SETTM zur Bestimmung der Rechenzeit für die verschiedenen Schritte.

Für die Berechnung der pseudoinversen Matrix nach der Methode (I.9.d) wurde für das M.G.S.-Verfahren (Modifizierte Gram-Schmidt-Orthogonalisierungsverfahren) das FORTRAN Programm von I. Barrodale und G.F. Stuart: A FORTRAN PROGRAM FOR LINEAR LEAST-SQUARES PROBLEMS OF VARIABLE DEGREE, Proceedings of the fourth Manitoba Conference on Numerical Mathematics, October 2-5 1974, pages 191-204, verwendet.

Die Resultate für die Beispiele $a, b, c_{1,2}, d_{1,2,3}, e_{1,2}$ sind in den Tabellen I, II, III zusammengestellt.

Verfahren	Anzahl Iterationen für Beispiele				Referenzabweichung für Beispiele			
	a	b	c ₁	c ₂	a	b	c ₁	c ₂
CHENEY [8]	2	4	4	4	1.555554	2.624498	1.000097	1.000038
BARRODALE [1]	5	4	3	3	1.555555	2.625000	1.000099	1.000038
VORGESCHLAGENES	2	2	2	2	1.555555	2.624999	1.000093	0.999997

Tabelle I

Verfahren	Anzahl Iterationen für Beispiele			Referenzabweichung für Beispiele		
	d ₁	d ₂	d ₃	d ₁	d ₂	d ₃
CHENEY [8]	13	13	13	0.0654689	0.0289735	0.00981553
BARRODALE [1]	11	16	21	0.0654678	0.0296628	0.00881875
VORGESCHLAGENES	5	8	8	0.0654678	0.0296628	0.00881875

Tabelle II

Verfahren	Anzahl Iterationen für Beispiele		Referenzabweichung für Beispiele	
	e_1	e_2	e_1	e_2
CHENEY [8]	*	*	*	*
BARRODALE [1]	20	26	0.00899203	0.00899203
VORGESCHLAGENES	13	13	0.00899203	0.00899203

Tabelle III

* Der Algorithmus vom Cheney [8] hat für die Beispiele e_1 , e_2 keine richtige Lösung geliefert, weil die inverse Matrix jeder neuen Referenz nach der Formel (IV.10.1) bestimmt wurde und diese Formel zu einer numerischen Instabilität geführt hat

$$c_{i\beta}^k = \frac{1}{y_\beta} c_{i\beta}^{k-1} \quad (\text{IV.10.1})$$

$$c_{ij}^k = c_{ij}^{k-1} - c_{i\beta}^k y_j \quad j \neq \beta$$

(c_{ij}^k das ij -te Element der inversen Matrix beim k -ten Schritt).

Unter Benützung von Gleitpunktarithmetik ergibt sich das berechnete c_{ij}^k wie folgt:

$$c_{ij}^k = \text{gl}(c_{ij}^{k-1} - \frac{y_j}{y_\beta} c_{i\beta}^{k-1}) = (c_{ij}^{k-1} - \frac{y_j}{y_\beta} c_{i\beta}^{k-1}(1+\epsilon_1))(1+\epsilon_2)$$

$$\epsilon_{ij} = c_{ij}^k - (c_{ij}^{k-1} - \frac{y_j}{y_\beta} c_{i\beta}^{k-1}) = c_{ij}^k - (\frac{c_{ij}^k}{1+\epsilon_2} + \frac{y_j}{y_\beta} c_{i\beta}^{k-1}) \epsilon_1 =$$

$$= \frac{\epsilon_2}{1+\epsilon_2} c_{ij}^k + \frac{y_j}{y_\beta} c_{i\beta}^{k-1} \epsilon_1$$

Weil c_{ij}^k auch gleich Null sein kann und die Faktoren $\frac{y_j}{y_\beta}$

gross werden können, können keine Schranken für ϵ_{ij} angegeben werden.

Um die Stabilität der vorgeschlagenen Aenderungsmethode für die pseudoinverse Matrix zu prüfen, berechnet man die singulären Werte der Matrix A_I der ersten Referenz $[I]$, sowie die singulären Werte der Matrix $A_{I'}$ der Endreferenz $[I']$, ferner die pseudoinverse Matrix $A_{I_L}^+$ der ersten Referenz, bzw. $A_{I_G}^+$ nach der Methode (I.9.b), bzw. (I.9.d).

Nach der Durchführung des Algorithmus werden die pseudoinversen Matrizen $A_{I_L}^+$, bzw. $A_{I_G}^+$ der Endreferenz $[I']$ nach der Methode (I.9.b), bzw. (I.9.d) berechnet.

Die durch den vorgeschlagenen Algorithmus erhaltene Matrix $A_{I_V}^+$ der

Endreferenz vergleicht man mit den Matrizen $A_{I_L}^+$, $A_{I_G}^+$, indem man

die minimale Anzahl der übereinstimmenden Dezimalstellen angibt.

Die Resultate sind in der Tabelle IV ersichtlich. Die zweite und dritte Kolonne beziehen sich auf das gleiche Beispiel e_2 , aber mit verschiedenen Startreferenzen.

Der vorgeschlagene Algorithmus hat für alle Beispiele weniger Iterationen benötigt und für die Beispiele e_1 , bzw. e_2 78%, bzw. 81% der Rechenzeit des Algorithmus von Barrodale und Phillips [1] beansprucht.

Beispiele	e_1	e_2	e_2
Singulare Werte der Matrix A_I (Anfangsreferenz)	$0.41587836437169 \times 10^1 > \dots$ $> 0.1709280822933 \times 10^{-7}$	$0.394602222600973 \times 10^1 > \dots$ $> 0.22846616710642 \times 10^{-2}$	$0.39664034323097 \times 10^1 > \dots$ $> 0.16260921232855 \times 10^1$
Anzahl Übereinstimmender Stellen der Matrizen $A_{I_G}^+$, $A_{I_L}^+$	4	12	15
Singulare Werte der Matrix A_I (Endreferenz)	$0.54990376622474 \times 10^1 > \dots$ $> 0.27639791182166 \times 10^{-6}$	$0.38357921694214 \times 10^1 > \dots$ $> 0.20442248845729 \times 10^1$	$0.38357921694214 \times 10^1 > \dots$ $> 0.20442248845729 \times 10^1$
Anzahl Übereinstimmender Stellen der Matrizen $A_{I_G}^+$, $A_{I_V}^+$	8	14	15
Anzahl Übereinstimmender Stellen der Matrizen $A_{I_L}^+$, $A_{I_V}^+$	8	14	15

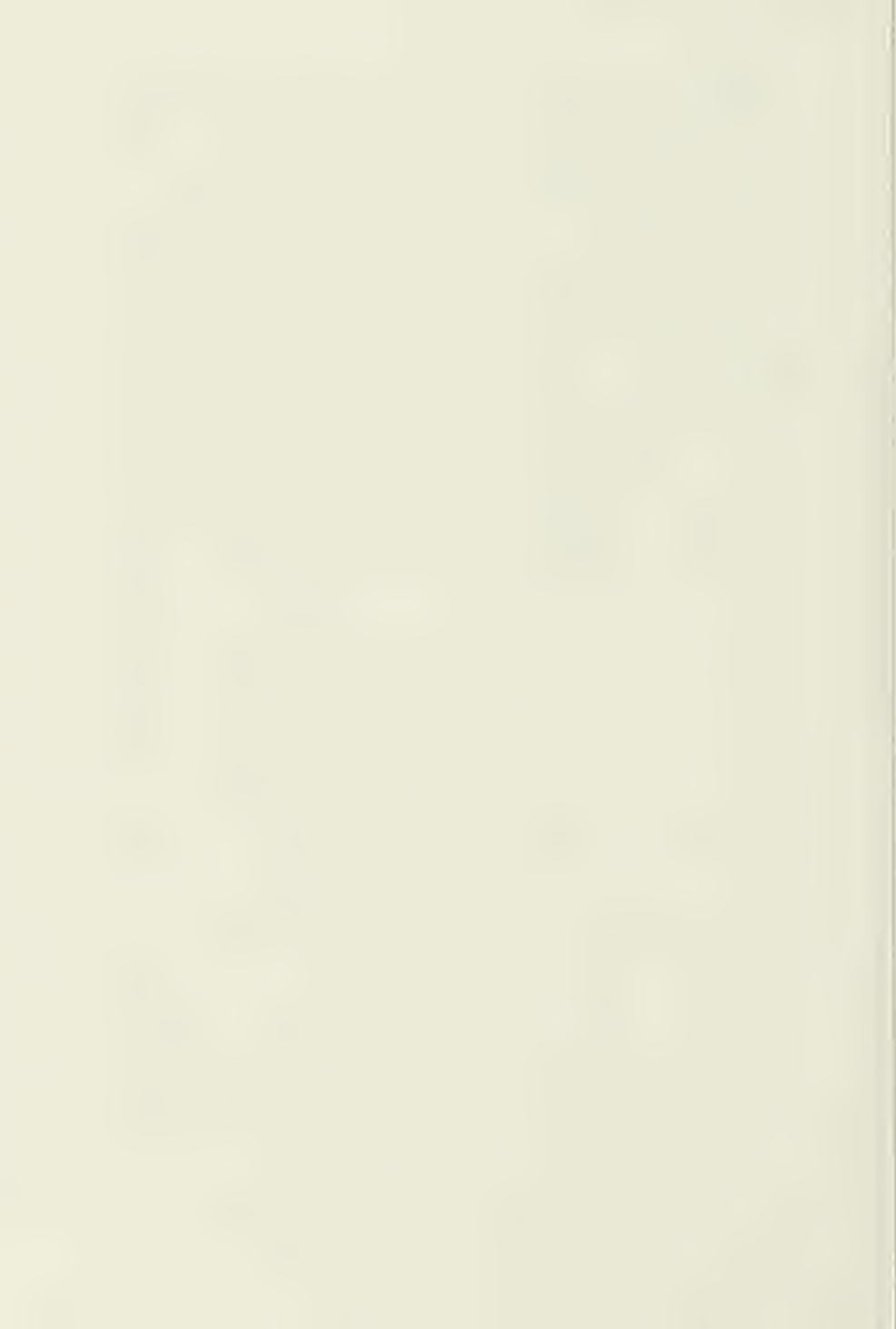
Tabelle IV

LITERATUR

- [1] BARRODALE I.-PHILLIPS C. : An improved algorithm for discrete Chebyshev linear approximation. Proceedings of fourth Manitoba Conference on Numerical Mathematics. October 2-5, 1974.
- [2] BARRODALE I.-STUART G.F. : A FORTRAN programm for linear least-squares problems of variable degree. Proceedings of the fourth Manitoba Conference on Numerical Mathematics. October 2-5, 1974.
- [3] BARTELS R.H.-GOLUB G.H. : Stable numerical methods for obtaining the Chebyshev solution to an overdetermined system of equations. Communications of the ACM V. 11,1968 (401-406).
- [4] BITTNER L. : Das Austauschverfahren der linearen Tschebyscheff-Approximation bei nicht erfüllter Haarscher Bedingung. ZAMM 41,1961 (238-256).
- [5] BJÖRK ÄKE : Solving linear least squares problems by Gramm-Schmidt orthogonalization. BIT 7, 1967 (1-21).
- [6] CHENEY E.W. : Introduction to approximation theory. McGraw-Hill, 1966.
- [7] CHENEY E.W. : Best approximation on convex sets in a metric space. Journal of approximation theory 12,1974 (94-97).
- [8] CHENEY E.W.-GOLDSTEIN A.A. : Newton's method for convex programming and Tchebycheff approximation. Numerische Mathematik 1,1959 (253-268).
- [9] CHENEY E.W.-GOLDSTEIN A.A. : A finite algorithm for the solution of consistent linear equations and inequalities and for the Tchebycheff approximation of inconsistent linear equations. Pacific J. Math. 8,1958 (415-427).
- [10] DURIS C.S.-SREEDHARAN V.P. : Chebyshev and l^1 - solutions of linear equations using least squares solutions. SIAM J. Num. Anal. 5,1968 (491-505).
- [11] EASTON M.C. : A fixed point method for Tchebycheff solution of inconsistent linear equations. J. Inst. Maths Applic. 12, 1973 (137-159).
- [12] GOLUB G.-REINSCH C. : Singular value decomposition and least squares solutions. Numer. Math. 14,1970 (403-420).

- [13] IVANOV V.V. : A general approximation method for solving linear problems. Soviet. Math. Dokl. 3, 1962 (415-418).
- [14] KUHNERT F. : Pseudoinverse Matrizen und die Methode der Regularisierung. Teubner-Texte zur Mathematik, Leipzig 1976.
- [15] LAWSON C.L.-HANSON R.J. : Solving least squares problems. Prentice-Hall, Inc. 1974.
- [16] LEVINE N.-GOLDSTEIN A.A.-HERESHOFF J.B. : On the "Best" and "Least Qth" approximation of an overdetermined system of linear equations. J. Assoc. Comput. Mach. 4, 1957 (341-347)
- [17] MEICLER M. : Chebyshev solution of an inconsistent system of $n+1$ linear equations in n unknowns in terms of its least squares solution. SIAM Review 10, 1968 (373-375).
- [18] MOURSUND D. : Chebyshev solutions of $n+1$ linear equations in n unknowns. J. Ass. Comput. Mach. 12, 1965 (382-387).
- [19] MUELLER H.N. : Zur Eindeutigkeit bei der linearen Approximation im allgemeinen sowie Tschebyscheffschen Fall. Numer. Math. 11, 1968 (144-150).
- [20] OSBORNE M.R.-WATSON G.A. : On the best linear Chebyshev approximation. Comput. J. 10, 1967 (172-177).
- [21] OSBORNE E.E. : Smallest least squares solution of linear equation. J. SIAM Numer. Anal. Ser. B 2, 1965 (300-307).
- [22] POWELL M.J.D.-HOPPER M.J. : A technique that gains speed and accuracy in the minimax solution of overdetermined linear equations. Mathematical Software III, Edited by J.R. RICE 1977.
- [23] RIVLIN T.J. : Overdetermined system of linear equations. SIAM Review 5, 1963 (52-66).
- [24] ROSEN J.B. : Chebyshev solution of large linear systems. J. of Comp. and System Sciences 1, 1967 (29-43).
- [25] RUST B.-BURRUS W.R.-SCHNEEBERGER C. : A simple algorithm for computing the generalized inverse of a matrix. Comm. of the ACM 9, 1966 (381-387).
- [26] SAUTTER W. : Fehlerfortpflanzung und Rundungsfehler bei der verallgemeinerten Inversion von Matrizen. Dissertation, TU München, 1971.
- [27] SCHWARZ H.R. : Matrizen Numerik. 2. Auflage, Stuttgart 1972.

- [28] SHINOZAKI N.-SIBUYA M.-TANABE K. : Numerical algorithms for the Moore-Penrose inverse of a matrix: Direct methods. Ann. Inst. Statist. Math., 24, 1971 (193-203).
- [29] STIEFEL E. : Ueber diskrete und lineare Tschebyscheff-Approximationen. Num. Math. 1, 1959 (1-28).
- [30] STIEFEL E. : Note on Jordan elimination, linear programming and Tchebycheff approximation. Numr. Math. 2, 1960 (1-17).
- [31] WILKINSON J.H.-PETERS G. : The least squares problem and pseudoinverses. Comp. J. 13, 1970 (309-316).
- [32] WILKINSON J.H. : Rundungsfehler. Springer-Verlag 1969.
- [33] ZUKHOVITSKIJ S.I.-POLJAK R.A.-PRIMAK M.E. : An algorithm for solving the convex Tchebyshev approximation problem. Soviet Math. Dokl. 4, 1963 (901-904).
- [34] ZIELKE G. : Numerische Berechnung von benachbarten inversen Matrizen und linearen Gleichungssystemen. Schriften zur Datenverarbeitung. Band 2, 1970.



Lebenslauf

Am 27. September 1944 wurde ich, Eteoklis Rekkas, in Athen (Griechenland) als Sohn des Berufsoffiziers Sotirios Rekkas und seiner Frau Athanasia geb. Pipilaidou geboren. Ich besuchte das Gymnasium und die Universität in Athen. Im Frühling 1968 erwarb ich dort das Diplom für Mathematik mit der Note "sehr gut". Im Oktober 1968 reiste ich in die Schweiz, wo ich im Sommersemester 1969 an der Philosophischen Fakultät II der Universität Zürich ein Nachdiplomstudium begann. Nach zwei Semestern wurde mein Mathematikdiplom aus Athen von der Philosophischen Fakultät II der Universität Zürich anerkannt. Bis Ende 1972 war ich neben dem Nachdiplomstudium mathematischer Mitarbeiter bei den Projekten "Die Stellung der Frau in der Schweiz" und "Industrial structures with a simulation model of some behavioral characteristics of large firms" des Soziologischen Instituts der Universität Zürich.

Im Januar 1973 wurde ich zum obligatorischen Militärdienst nach Griechenland eingezogen und wurde vom Januar 1974 bis März 1975 in der EDV-und Operations Research-Abteilung der Griechischen Armee als Mathematiker eingesetzt. Die Arbeit an der vorliegenden Dissertation begann ich im Wintersemester 1975/1976, wobei auf Empfehlung von Herrn Prof. Dr. H.P. Künzi ab Frühling 1978 Herr Prof. Dr. H.R. Schwarz die Betreuung übernahm. Während meines Studiums an der Universität Zürich besuchte ich Vorlesungen und Kurse folgender Dozenten: K. Bauknecht, H. Gross, P. Kall, H. Künzi, W. Oettli, H.R. Schwarz, B. van der Waerden, F. Weinberg.

